Un système de visualisation pour l'extraction, l'évaluation, et l'exploration interactives des règles d'association.

Julien Blanchard

École Centrale de Nantes **Université de Nantes** **École des Mines de Nantes**

ÉCOLE DOCTORALE S T I M

« SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DES MATÉRIAUX »

Année 2005 N° ED 366-220

THÈSE DE DOCTORAT

Spécialité : INFORMATIQUE

présentée et soutenue publiquement par

Julien BLANCHARD

le 24 novembre 2005
à l'École Polytechnique de l'Université de Nantes

Un système de visualisation pour l'extraction, l'évaluation, et l'exploration interactives des règles d'association

Président :	Gilbert RITSCHARD, Professeur à l'Université de Genève
Rapporteurs :	Jean-Marie PINON, Professeur à l'INSA de Lyon
	Gilles VENTURINI, Professeur à l'École Polytechnique de l'Université de Tours
Examinateurs :	Bernard DOUSSET, Professeur à l'Université Paul Sabatier
	Henri BRIAND, Professeur à l'École Polytechnique de l'Université de Nantes
	Fabrice GUILLET, Maître de conférences à l'École Polytechnique de l'Université de Nantes
Invités :	Régis GRAS, Professeur émérite à l'École Polytechnique de l'Université de Nantes
	Jacques PHILIPPÉ, Expert scientifique à PerformanSe Corp.

Directeur de thèse : Henri BRIAND
Laboratoire : Laboratoire d'Informatique de Nantes Atlantique (LINA)
2, rue de la Houssinière – BP 92208 – 44322 Nantes Cedex 3

À ma femme.

Je tiens à remercier Henri Briand qui m'a accueilli dans son équipe et a dirigé mes travaux de thèse. Il a su me montrer la voie à suivre et me faire profiter de son expérience de la recherche tout en m'accordant une grande liberté de travail. Je remercie mon encadrant Fabrice Guillet pour son soutien et ses conseils au quotidien. Il m'a aidé à développer mes aptitudes pour la recherche et l'enseignement, je lui en suis très reconnaissant. Je remercie Régis Gras, dont les travaux ont significativement influencé cette thèse. Dès l'année de mon DEA, nos discussions et débats m'ont particulièrement stimulé. J'ai également apprécié sa gentillesse et son enthousiasme. Je remercie Pascale Kuntz pour ses judicieux conseils ainsi que son soutien dynamique dans les moments difficiles. Je remercie Jacques Philippé qui m'a permis d'obtenir une bourse de la *fondation VediorBis pour la Recherche et l'Emploi* afin de financer ma thèse. Je remercie MM. Jean-Marie Pinon et Gilles Venturini d'avoir accepté d'être les rapporteurs de ma thèse, et pour l'attention avec laquelle ils l'ont lue et évaluée. Je remercie MM. Gilbert Ritschard et Bernard Dousset d'avoir accepté de faire partie de mon jury.

Je remercie Catherine Galais pour ses relectures minutieuses des articles anglophones. Je remercie toutes les personnes de l'équipe COD qui m'ont soutenu ou aidé, et notamment Rémi Lehn, Bruno Pinaud, et Jérôme David. Enfin, je remercie Aline pour son dévouement, sa patience, et sa confiance tout au long de ces quatre années.

Table des matières

Liste des figures

Liste des tableaux

Liste des Algorithmes

Introduction

Extraction de Connaissances dans les Données

Depuis l'apparition des bases de données dans les années 60, l'augmentation des capacités des solutions de sauvegarde fait exploser le volume de données stockées sous forme numérique dans le monde. On estime souvent qu'il double tous les vingt mois [Kod97]. Cette évolution exponentielle a été confirmée récemment dans une étude de l'université de Californie de Berkeley qui fait état d'une augmentation de 114% entre 1999 et 2002 de la quantité d'information produite annuellement sur disque dur dans le monde [LV03]. Dans les années 90, cette accumulation d'informations dans les bases de données a motivé le développement d'un nouveau champ de recherche : l'Extraction de Connaissances dans les bases de Données (ECD). L'ECD consiste à mettre en évidence des connaissances nouvelles, valides, et potentiellement utiles dans de grandes bases de données [FPSM91]. Ce domaine d'étude emprunte à la fois à la statistique, à l'analyse de données, et à l'intelligence artificielle [HMS01] [HK00]. D'un point de vue général, les spécificités des applications de l'ECD découlent du fait que les données étudiées proviennent de bases de données :

- les volumes de données sont grands, voire colossaux[1], et résident non pas en mémoire centrale mais en mémoire secondaire (disque, bande) ;
- conséquemment, et dans un souci d'efficacité, les applications de l'ECD sont fortement couplées aux systèmes de gestion de bases de données ;
- dans le cas général, les données en entrée d'une application d'ECD sont des données brutes d'exploitation, avec tout ce que cela implique (données redondantes, erronées, incomplètes, réparties, dynamiques, en flux, etc.), et non pas des données collectées et préparées en vue d'une analyse.

De plus, l'ECD vise à produire des résultats qui s'adressent directement à des utilisateurs métier (experts des données étudiées), et non à des statisticiens. L'intelligibilité des résultats relève donc d'une attention toute particulière en ECD.

[1] La taille d'une base de données se mesure aujourd'hui en gigaoctets pour les bases moyennes (2^{30} octets soit environ un milliard d'octets), et en téraoctets pour les bases les plus grandes (2^{40} octets soit environ mille milliards d'octets). D'après l'étude de référence de Winter Corporation (*www.wintercorp.com*), la plus grande base de données d'entreprise du monde est en 2005 l'entrepôt de données de Yahoo Search Marketing (référencement payant sur Internet), avec 100 téraoctets pour 385 milliards d'enregistrements. L'entrepôt de données de France Télécom était en tête du classement dans l'édition précédente de l'étude (2003).

Le processus d'ECD se déroule en trois étapes [FPSS96] :

- le pré-traitement, qui consiste à "nettoyer" et à mettre en forme les données[2] (sélection des données, élimination des doublons, élimination des valeurs aberrantes, gestion des valeurs manquantes, transformation des variables, création de nouvelles variables, etc.) ;
- la fouille de données (*data mining*), étape moteur de l'ECD qui consiste à identifier les motifs qui structurent les données, ou produire des modèles explicatifs ou prédictifs des données ;
- le post-traitement, qui consiste à mettre en forme et évaluer les résultats obtenus (appelés connaissances), et à les faire interpréter et valider par l'utilisateur.

Ce processus est itératif et hautement interactif [FPSS96] [ST96a]. L'utilisateur y est impliqué à chaque étape, tant pour effectuer des choix (quels prétraitements ? quels paramètres pour l'algorithme de fouille de données ? quels post-traitements ?) que pour examiner si nécessaire les données, les résultats intermédiaires, ou les connaissances produites.

Règles d'association

En sciences cognitives, de nombreuses théories de représentation de la connaissance sont fondées sur les règles [HHNT86]. D'une manière générale, les règles sont des propositions de la forme "si *prémisse* alors *conclusion*", notées *prémisse* $\rightarrow$ *conclusion*. Elles ont l'avantage de représenter les connaissances de manière explicite (contrairement aux modèles connectionnistes par exemple), et sont d'ailleurs le modèle prépondérant de nombreuses applications d'intelligence artificielle, en particulier les systèmes experts. En ECD, une des principales méthodes produisant des connaissances sous forme de règles est *l'extraction de règles d'association*, introduite par Agrawal, Imielienski et Swami [AIS93]. Etant donnée une table dans une base de données relationnelle, les règles d'association sont des tendances implicatives *prémisse* $\rightarrow$ *conclusion* où la prémisse et la conclusion sont des expressions qui portent sur les attributs de la table et indiquent les valeurs qu'ils doivent prendre. Ces règles signifient que si un enregistrement de la table vérifie la prémisse, alors il vérifie sûrement également la conclusion. L'une des premières applications des règles d'association, et sans doute la plus connue, a été l'étude du panier de la ménagère. Elle consiste à découvrir des combinaisons de produits qui sont souvent achetés ensemble dans un supermarché, du type "si un client achète des huîtres, alors il achète sûrement aussi du muscadet".

Depuis l'algorithme de référence d'Agrawal et Srikant [AS94a], nommé *Apriori*, de nombreux algorithmes ont été développés pour extraire efficacement des règles d'association (voir [HGN00] pour une synthèse). Ces algorithmes valident les règles avec deux mesures : le support et la confiance. Le support est la proportion d'enregistrements de la table qui vérifient la prémisse et la conclusion (par exemple, 1% des clients achètent des huîtres et du muscadet), tandis que la confiance est la proportion d'enregistrements qui vérifient la conclusion parmi

[2] Dans le cas où les données proviennent non pas d'une base d'exploitation mais d'un entrepôt de données, une partie du pré-traitement est déjà réalisée.

ceux qui vérifient la prémisse (par exemple, 90% des clients qui achètent des huîtres achètent du muscadet). Les algorithmes de découverte de règles d'association exécutent tous la même tâche déterministe : étant donnés un seuil minimal de support et un seuil minimal de confiance, produire l'ensemble exhaustif de toutes les règles qui possèdent un support et une confiance supérieurs aux seuils. Ces algorithmes ont la particularité d'être non supervisés, c'est-à-dire qu'ils ne nécessitent pas qu'on leur précise des attributs endogènes mais au contraire envisagent toutes les combinaisons possibles d'attributs pour la prémisse et pour la conclusion. Cette nature non supervisée fait la force des règles d'association : les algorithmes ne requièrent aucune connaissance préalable sur les données de la part de l'utilisateur (idéal pour débuter une étude sur des données), et surtout ils peuvent découvrir des règles que l'utilisateur juge intéressantes alors même qu'elles sont constituées de combinaisons d'attributs auxquelles il n'aurait pas nécessairement songé. Cependant, la nature non supervisée de la découverte de règles d'association constitue également la principale limite de la méthode. La quantité de règles générée par un algorithme croît en effet exponentiellement avec le nombre d'attributs décrivant les données. Dans la pratique, les volumes de règles obtenus sont prohibitifs, atteignant rapidement plusieurs centaines de milliers de règles.

Post-traitement des règles : comment faire face au volume de règles ?

Du fait des grandes quantités de règles que produisent les algorithmes de fouille de données, le post-traitement est une étape nécessaire mais difficile dans un processus de recherche de règles d'association. Il consiste en une seconde opération de fouille, mais alors que la fouille de données est réalisée automatiquement par des algorithmes combinatoires, la fouille de règles est généralement laissée à la charge de l'utilisateur. En pratique, il est très laborieux pour ce dernier de rechercher des connaissances intéressantes dans les listes de règles obtenues à la sortie des algorithmes.

Différentes solutions ont été proposées pour aider l'utilisateur dans sa tâche.

- De nombreux indices de qualité ont été développés afin d'évaluer les règles selon différents points de vue [BSGG04] [Fre98] [BMS97] [PS91]. Ils permettent à l'utilisateur d'identifier et de rejeter les règles de faible qualité, mais aussi d'ordonner les règles acceptables des meilleures aux plus mauvaises.
- Une autre solution consiste à organiser une exploration interactive des règles pour l'utilisateur [FR04] [TA02] [MLW00] [KMR$^+$94]. Plusieurs logiciels et langages de requêtes ont été conçus dans cette optique.
- La tâche de l'utilisateur peut également être facilitée en lui soumettant des représentations visuelles des règles [HHC03] [HW01] [Leh00] [WWT99]. Elles facilitent la compréhension et accélèrent l'appropriation des règles par l'utilisateur.

Malgré ces différents travaux, plusieurs problèmes demeurent. Tout d'abord, les indices de qualité sont nombreux et souvent redondants entre eux [LMV$^+$04]

[TKS04] [VLL04]. La signification de ces indices n'est pas non plus très claire fréquemment pour l'utilisateur. D'une manière générale, il est difficile de choisir quels indices appliquer.

Ensuite, l'interactivité dans le post-traitement de règles d'association est souvent pauvre : les interactions ne sont pas pleinement adaptées à la tâche de l'utilisateur, et en particulier elles ne tiennent pas compte de la spécificité des données, c'est-à-dire le fait qu'il s'agisse de règles. Pour mieux prendre en considération les efforts cognitifs de l'utilisateur, le processus d'ECD doit être considéré non pas sous l'angle des algorithmes de fouille de données mais sous l'angle de l'utilisateur, comme un système d'aide à la décision centré sur l'utilisateur [BA96].

Enfin, les représentations visuelles de règles sont généralement peu interactives, voire même statiques. Elles sont utilisées comme un outil complémentaire, pour présenter les résultats sous une forme plus compréhensible, mais ne permettent pas de fouiller les ensembles de règles. De plus, les indices de qualité intégrés dans ces représentations sont peu nombreux (trois au maximum, souvent deux) et faiblement mis en valeur, alors que ce sont des indicateurs essentiels pour le post-traitement des règles.

Apports de la visualisation

La *visualisation d'information* [CMS99] [Spe00] consiste à représenter des informations abstraites sous forme visuelle afin d'améliorer la cognition pour une tâche donnée (c'est-à-dire l'acquisition et l'utilisation de nouvelles connaissances). Contrairement à la visualisation scientifique qui s'attache à représenter des entités réelles, la visualisation d'information porte sur des informations abstraites. Il s'agit donc de donner une représentation visuelle à des informations qui n'induisent aucune représentation évidente. Si la visualisation d'information est un champ d'étude large et ancien qui trouve ses origines dans la cartographie et les graphiques statistiques, elle constitue aujourd'hui une discipline à part entière, qui emprunte également à la psychologie cognitive et à la sémiotique, ainsi qu'aux interfaces homme-machine et à l'imagerie de synthèse.

La visualisation améliore la cognition grâce aux capacités perceptives du système visuel humain. Sans entrer dans des considérations relevant de la psychologie cognitive ou de la neurophysiologie, on peut dire que d'une manière générale, la visualisation [CMS99] [War00] [CRC03] :

- facilite l'identification de ressemblances,
- facilite l'identification de structures,
- facilite l'identification de singularités,
- facilite la mémorisation,
- oriente la réflexion de l'utilisateur et facilite la génération d'hypothèses,
- et tout ceci sur des données qui peuvent être très volumineuses.

En particulier, certaines informations visuelles sont traitées de manière inconsciente et très rapide par le cerveau, sans même réduire sa disponibilité pour d'autres tâches (on parle de perception pré-attentive). C'est le cas par exemple de la position, de la taille, ou de la couleur [Ber67] [CMS99]. Ainsi, il est instan-

tané de trouver sur une carte météorologique le temps qu'il fera le lendemain dans sa ville, ou bien de déterminer parmi les trente dernières années celle qui a été la plus pluvieuse sur un histogramme pluviométrique. Exécuter les mêmes tâches à partir d'informations textuelles est en revanche une activité consciente, qui nécessite plus de temps, et qui consomme une partie des "ressources" du cerveau.

Avec l'avènement de l'informatique, la visualisation est devenue dynamique, il s'agit d'une activité interactive. Dans sa théorie écologique de la perception [Gib79], Gibson établit que la perception est indissociable de l'action : il faut agir pour percevoir et percevoir pour agir [HBL01]. Ainsi, la visualisation d'information étudie non seulement les meilleures représentations à produire pour améliorer la cognition, mais aussi les meilleures interactions à mettre en oeuvre sur ces représentations [CMS99] [Spe00].

Contributions de la thèse

Les contributions de la thèse se déclinent en quatre thèmes. Tout d'abord, nous étudions deux solutions pour assister l'utilisateur dans le post-traitement des règles d'association : la mesure de la qualité des règles, et la visualisation interactive des règles (la seconde approche exploite la première). Nous proposons ensuite d'adapter l'extraction des règles au caractère interactif du post-traitement en développant des algorithmes spécifiques pour l'extraction locale des règles. Enfin, ces trois approches sont mises en oeuvre au sein de l'outil de visualisation *ARVis* (*Association Rule Visualization*).

1. Mesure de la qualité des règles
Nous formalisons les notions de *règle* et d'*indice de règle*, puis réalisons une classification inédite des nombreux indices de la littérature. En clarifiant leur signification, une telle classification permet d'aider l'utilisateur à choisir les indices pertinents pour son besoin. Nous proposons également de nouveaux indices aux propriétés originales : l'indice probabiliste d'écart à l'équilibre, l'intensité d'implication entropique, et le taux informationnel.

2. Visualisation interactive des règles
Nous établissons une méthodologie pour la visualisation interactive des règles d'association, nommée *Rule Focusing*. Elle est conçue pour faciliter la tâche de l'utilisateur confronté à de grands ensembles de règles en prenant en compte ses capacités de traitement de l'information. Dans cette méthodologie, l'utilisateur explore par lui-même des petits ensembles successifs de règles au moyen d'une visualisation interactive pourvue d'opérateurs de navigation adaptés. Cette approche est fondée sur :
- des principes de visualisation d'information pour la construction de représentations efficaces [Ber67], et plus particulièrement pour la mise en valeur des indices de qualité ;
- des principes cognitifs de traitement de l'information dans le contexte des modèles de décision [Mon83].

3. **Extraction locale des règles**
   Nous développons des algorithmes spécifiques pour l'extraction locale des règles, qui permettent de n'extraire que les ensembles de règles que l'utilisateur souhaite visualiser. Ces algorithmes exploitent des contraintes puissantes qui restreignent drastiquement l'espace de recherche. Ils donnent la possibilité de s'affranchir des limites des algorithmes d'extraction exhaustifs comme *Apriori*. Ainsi, en explorant les règles, l'utilisateur dirige à la fois l'extraction et le post-traitement des connaissances.

4. **Outil de visualisation 3D *ARVis***
   ARVis est un outil opérationnel pour la visualisation interactive des règles d'association, qui met en oeuvre les trois approches précédentes. Il permet d'explorer de grands volumes de règles et d'identifier les connaissances pertinentes. L'outil repose sur une visualisation intuitive en trois dimensions qui supporte de grands ensembles de règles décrits par plusieurs indices.

Organisation de la thèse

Cette thèse est constituée de trois parties comprenant chacune deux chapitres. La première partie étudie les règles en tant qu'entité statistique purement conceptuelle, c'est-à-dire indépendamment des algorithmes de fouille de données utilisés pour les générer (il peut s'agir autant de règles d'association que de règles de classification issues d'arbres de décision ou d'algorithmes d'induction). Cette partie est plus particulièrement consacrée à l'évaluation de la qualité des règles. Dans le **chapitre 1**, nous définissons les notions de *règle* et d'*indice de règle* et proposons notre classification des trente principaux indices. Dans le **chapitre 2**, nous présentons nos nouveaux indices : l'indice probabiliste d'écart à l'équilibre, l'intensité d'implication entropique, et le taux informationnel. Pour chaque indice, nous décrivons sa construction et étudions ses propriétés.

La deuxième partie est consacrée à l'extraction et au post-traitement des règles d'association. Dans le **chapitre 3**, nous nous intéressons aux deux principales catégories d'algorithmes d'extraction de règles d'association : les algorithmes exhaustifs et les algorithmes à contraintes. Nous passons en revue leurs caractéristiques puis comparons les deux approches. Le **chapitre 4** concerne la méthodologie *Rule Focusing* pour la visualisation interactive des règles d'association. Nous y réalisons un état de l'art sur le post-traitement des règles d'association, puis étudions deux tendances récentes de la visualisation d'information : les représentations 3D et la réalité virtuelle. Pour développer notre méthodologie, nous nous référons à des principes de visualisation d'information et des principes cognitifs de traitement de l'information.

La troisième partie est dédiée à l'outil de visualisation *ARVis*. Dans le **chapitre 5**, nous présentons les fonctionnalités des deux versions d'*ARVis* qui ont été réalisées, puis détaillons leur implémentation, et enfin exposons quelques exemples d'utilisation. Dans le **chapitre 6**, nous décrivons les algorithmes mis en oeuvre dans *ARVis* pour l'extraction locale des règles, puis étudions les temps de réponse.

Première partie

Qualité des règles

1 Règles et mesures de qualité

Sommaire

Les mesures de qualité de règles aident l'utilisateur à trouver des connaissances intéressantes dans les grands volumes de règles produits par les algorithmes de fouille de données. Ces mesures permettent :

- d'évaluer les règles selon différents points de vue,
- de rejeter celles qui sont trop mauvaises (utilisation d'un seuil de qualité minimale),
- d'ordonner celles qui sont acceptables des meilleures aux plus mauvaises.

Il existe deux catégories de mesures : les mesures subjectives (orientées utilisateur) et les mesures objectives (orientées données). Les mesures subjectives prennent en compte les objectifs de l'utilisateur et ses connaissances a priori sur les données [LHCM00] [PT99] [ST96b]. En revanche, seuls les cardinaux dus à

la contingence des données interviennent dans le calcul des mesures objectives. Parmi ces dernières, on trouve aussi bien des mesures fréquentielles rudimentaires que des mesures fondées sur des modèles probabilistes, en passant par des mesures issues de la théorie de l'information ou des mesures statistiques usuelles de liaison [Gui04]. Dans cette thèse, nous nous intéressons aux mesures objectives de qualité de règles, que nous appelons plus simplement *indices de règle*. En effet, la subjectivité du post-traitement de règles est prise en compte dans notre méthodologie d'exploration de règles non pas par l'intermédiaire de mesures subjectives de qualité mais au moyen d'outils interactifs (voir chapitre 4).

Après avoir introduit la terminologie et les notations utilisées dans le chapitre, nous présentons le concept de règle en section 1.2. La section 1.3 formalise la notion d'indice de règle et recense les principaux indices de la littérature. Enfin, nous réalisons en section 1.4 une classification inédite des indices de règle. En clarifiant leur signification, une telle classification aide l'utilisateur à choisir les indices pertinents pour son besoin.

1.1 Terminologie et notations

Nous considérons un ensemble E de n individus décrits par un ensemble V de variables qualitatives (il peut également s'agir de variables quantitatives discrétisées). En ECD, E est stocké sous forme de table dans une base de données relationnelle. Au moyen d'un codage disjonctif complet, l'ensemble V peut être remplacé par un ensemble I de variables booléennes de la forme *variable=modalité*. Ces nouvelles variables sont appelées des *items* dans la terminologie des règles d'association[1]. Par exemple, à partir de la variable *couleur_des_yeux* qui possède les trois modalités *marron*, *bleu*, et *vert*, on construit trois items : *couleur_des_yeux=marron*, *couleur_des_yeux=bleu*, et *couleur_des_yeux=vert*. Une conjonction d'items, comme par exemple (*couleur_des_yeux=marron* $\wedge$ *couleur_des_cheveux=brun*), est appelée un *itemset*. Il s'agit également d'une variable booléenne. La négation d'un itemset n'est pas un itemset puisque c'est une disjonction :

$$\overline{(couleur\_des\_yeux = marron \wedge couleur\_des\_cheveux = brun)} = \overline{(couleur\_des\_yeux = marron} \vee \overline{couleur\_des\_cheveux = brun)}$$

Cependant, il s'agit aussi d'une variable booléenne.

Soit a une variable booléenne qui est un itemset ou une négation d'itemset. La variable $\overline{a}$ est la négation de a. Nous notons A l'ensemble des individus de E qui vérifient a, et n_a le cardinal de A. Le complémentaire de A dans E est l'ensemble $\overline{A}$ de cardinal $n_{\overline{a}}$ (voir figure 1.1). La probabilité de l'événement a *est vrai* est notée $\mathrm{P}(a)$. Elle est estimée par la fréquence empirique (estimateur du maximum de vraisemblance) : $\mathrm{P}(a) = \frac{n_a}{n}$.

[1] Le mot "item" ("article" en français) est issu de l'étude du panier de la ménagère, application bien connue des règles d'association qui consiste à découvrir les combinaisons d'articles qui sont souvent achetés ensemble dans un supermarché. Dans cette application, les individus désignent un passage à la caisse. Ils sont appelés "transactions".

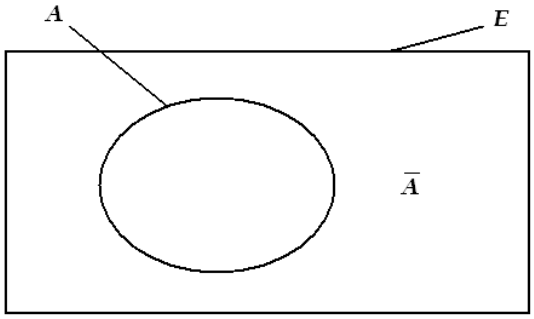


FIG. 1.1 – Diagramme de Venn de l'ensemble A dans la population E

a \ b	**1**	**0**	
1	n_{ab}	$n_{a\overline{b}}$	n_a
0	$n_{\overline{a}b}$	$n_{\overline{a}\overline{b}}$	$n_{\overline{a}}$
	n_b	$n_{\overline{b}}$	n

0 et 1 désignent respectivement la fausseté et la vérité.

TAB. 1.1 – Table de contingence croisant deux variables a et b

La ventilation des n individus de E selon deux variables booléennes a et b est donnée par la table de contingence croisant a et b (tableau 1.1), dans laquelle un effectif n_{ab} désigne le nombre d'individus vérifiant à la fois a et b. On a les relations suivantes entre les effectifs :

- $n_{ab} + n_{a\overline{b}} = n_a$,
- $n_{ab} + n_{\overline{a}b} = n_b$,
- $n_{\overline{a}b} + n_{\overline{a}\overline{b}} = n_{\overline{a}}$,
- $n_{a\overline{b}} + n_{\overline{a}\overline{b}} = n_{\overline{b}}$,
- $n_a + n_{\overline{a}} = n_b + n_{\overline{b}} = n$.

1.2 Règles

1.2.1 Liaisons entre variables qualitatives

Nous distinguons deux types de liaisons entre deux variables qualitatives nominales (voir tableau 1.2) :

- les liaisons entre variables (éventuellement) multimodales, qui traitent identiquement les différentes modalités de chaque variable ;
- les liaisons entre variables (nécessairement) binaires, qui ne traitent pas identiquement les deux modalités de chaque variable.

	Exemples de mesures de liaison	
	(symétriques)	**(orientées)**
liaison entre variables multimodales	χ^2, V de Cramer, T de Tschuprow, information mutuelle...	λ et τ de Goodman et Kruskal, coefficient d'incertitude de Theil...
liaison entre variables binaires	indice de Jaccard, indice de Dice, indice de Kulczynski, indice de Yule...	confiance, indice de Loevinger, intensité d'implication, conviction...

TAB. 1.2 – Différentes liaisons entre deux variables qualitatives nominales

Une liaison entre deux variables multimodales n'est pas affectée par la permutation des modalités d'une des variables (les mesures de liaison associées sont invariantes par cette permutation ; elles sont plutôt utilisées en classification supervisée où le modèle recherché doit expliquer toutes les modalités de la variable classe). En revanche, pour deux variables binaires v_1 et v_2, si l'on note $\overline{v_1}$ et $\overline{v_2}$ les variables issues respectivement de v_1 et v_2 par permutation des deux modalités ($\overline{v_1}$ et $\overline{v_2}$ sont les négations de v_1 et v_2 dans le cas booléen), alors la liaison entre v_1 et v_2 est différente de la liaison entre v_1 et $\overline{v_2}$ d'une part, et différente de la liaison entre $\overline{v_1}$ et v_2 d'autre part (les mesures de liaison entre variables binaires donnent des valeurs différentes).

Cette distinction entre les deux types de liaisons est importante car elle justifie que les mesures de liaison entre variables multimodales sont en général moins adaptées à l'évaluation des liaisons entre variables binaires. En effet, en traitant identiquement toutes les modalités, ces mesures ne distinguent pas vérité et fausseté, absence et présence, homme et femme, etc. Ainsi, elles évaluent identiquement les liaisons entre v_1 et v_2, entre v_1 et $\overline{v_2}$, et entre $\overline{v_1}$ et v_2. Cependant, si la liaison entre v_1 et v_2 est forte, alors intuitivement les liaisons entre v_1 et $\overline{v_2}$ et entre $\overline{v_1}$ et v_2 doivent être faibles. Ceci n'interdit pas de pouvoir utiliser des mesures entre variables multimodales sur des variables binaires, mais dans ce cas il faut nécessairement utiliser des mesures entre variables binaires en complément, afin de lever les symétries indésirables.

1.2.2 Définition d'une règle

Une règle est un cas particulier de liaison entre variables binaires : il s'agit d'une liaison orientée entre variables booléennes[2].

Définition 1.1 Une **règle** est un couple de variables booléennes (a, b) noté $a \rightarrow b$ où a et b sont des itemsets ou négations d'itemsets qui n'ont pas d'item

[2]Il est commun de considérer qu'une règle est implicitement une liaison de bonne qualité (si la règle est mauvaise pour les mesures de qualité utilisées, on dit intuitivement qu'il n'y a pas de règle). Dans cette partie de la thèse, étant donné que nous nous intéressons au problème de l'évaluation des règles, nous avons préféré définir une règle comme un simple couple de variables, indépendamment de toute considération sur la qualité, et dissocier la construction syntaxique d'une règle de son évaluation.

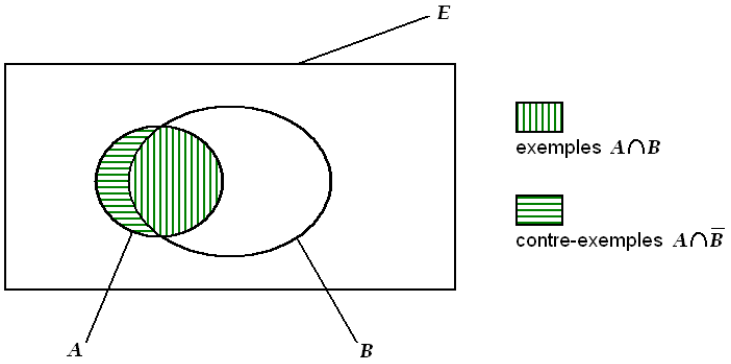


FIG. 1.2 – Diagramme de Venn pour la règle $a \rightarrow b$

en commun. Elle traduit la tendance de b à être vrai quand a est vrai, et peut se lire de la manière suivante : "si un individu vérifie a alors il vérifie sûrement b". a est la prémisse de la règle et b sa conclusion. Les exemples d'une règle sont les individus de $A \cap B$, c'est-à-dire ceux qui vérifient la prémisse et la conclusion, tandis que les contre-exemples sont les individus de $A \cap \overline{B}$, ceux qui vérifient la prémisse mais pas la conclusion (figure 1.2). Une règle est d'autant meilleure qu'elle admet beaucoup d'exemples et peu de contre-exemples.

Par conséquent, à partir de deux variables a et b il est possible de construire huit règles différentes :

- $a \rightarrow b$,
- $a \rightarrow \overline{b}$,
- $\overline{a} \rightarrow b$,
- $\overline{a} \rightarrow \overline{b}$,
- $b \rightarrow a$,
- $b \rightarrow \overline{a}$,
- $\overline{b} \rightarrow a$,
- $\overline{b} \rightarrow \overline{a}$.

Pour une règle $a \rightarrow b$, $a \rightarrow \overline{b}$ est la règle contraire, $b \rightarrow a$ est la règle réciproque, et $\overline{b} \rightarrow \overline{a}$ est la règle contraposée.

1.2.3 Support et confiance

Le support évalue la généralité d'une règle. Il s'agit de la proportion d'individus qui vérifient la règle dans le jeu de données [AIS93] :

$$support(a \rightarrow b) = \frac{n_{ab}}{n}$$

La confiance évalue la validité d'une règle. Il s'agit de la proportion d'individus qui vérifient la conclusion parmi ceux qui vérifient la prémisse [AIS93] :

$$confiance(a \rightarrow b) = \frac{n_{ab}}{n_a}$$

La confiance estime la probabilité conditionnelle que la variable b soit réalisée sachant que la variable a est réalisée. Elle peut aussi être interprétée comme le taux de réussite de la règle.

Le support et la confiance sont des indices simples, mais ils constituent les deux mesures les plus communément utilisées pour évaluer des règles. Tout d'abord parce qu'ils sont grandement intelligibles. Ensuite parce qu'ils sont à la base des algorithmes d'extraction de règles d'association (voir chapitre 3).

1.2.4 Cas particulier des règles logiques

Une règle logique est une règle qui n'admet aucun contre-exemple : $n_{a\bar{b}} = 0$. Sa confiance vaut donc 1. Une telle règle signifie "si un individu vérifie la prémisse, alors il vérifie systématiquement la conclusion". Ceci se traduit dans l'ensemble des individus par une inclusion : la règle logique $a \rightarrow b$ est équivalente à $A \subseteq B$ (figure 1.3). Une règle logique est aussi équivalente à sa règle contraposée.

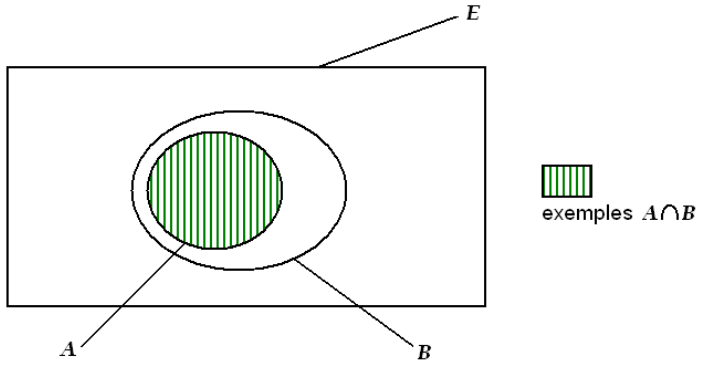


FIG. 1.3 – Diagramme de Venn pour la règle logique $a \rightarrow b$

La recherche de règles logiques dans des jeux de données a déjà été envisagée, par exemple dans [GD86] ou dans le système CHARADE [Gan91]. Cependant, une règle logique est extrêmement fragile puisque le moindre contre-exemple suffit à l'invalider, alors que les jeux de données réels sont largement bruités. De plus, les exceptions d'une règle peuvent s'avérer dignes d'intérêt pour l'utilisateur. En conséquence, on choisit en ECD de relaxer la contrainte logique et de tolérer des contre-exemples dans les règles.

1.2.5 Modélisation des règles

De même que la table de contingence de deux variables binaires est définie par la donnée de quatre effectifs indépendants, une règle peut être modélisée par quatre paramètres. Communément (comme le souligne Freitas [Fre99], il s'agit généralement dans la littérature d'une hypothèse implicite), on utilise comme paramètres de modélisation n_a, n_b, n et un effectif de la distribution jointe des deux variables comme n_{ab} ou $n_{a\bar{b}}$[3]. Comme Piatetsky-Shapiro [PS91],

[3]Certains auteurs font l'hypothèse que n est une constante, et ne considèrent donc que trois paramètres de modélisation. Cependant, ceci interdit de comparer des règles issues de bases de données différentes.

nous choisissons comme quatrième paramètre le nombre d'exemples n_{ab}. Chaque règle $a \rightarrow b$ est donc modélisée par $(n_{ab},\ n_a,\ n_b,\ n)$[4]. Dans la suite, nous ne distinguons pas une règle de son modèle : $(a \rightarrow b) = (n_{ab},\ n_a,\ n_b,\ n)$. L'ensemble $\mathcal{R}$ des règles possibles est le sous-ensemble de $\mathbb{N}^4$ suivant :

$$\mathcal{R} = \left\{ (n_{ab},\ n_a,\ n_b,\ n) \mid n_a \leq n,\ n_b \leq n,\ \max(0,\ n_a+n_b-n) \leq n_{ab} \leq \min(n_a,\ n_b) \right\}$$

Le choix des paramètres de modélisation est important puisque, même si tous les choix sont équivalents, il conditionne l'étude des liaisons entre variables en induisant un certain point de vue pour les simulations numériques des mesures. Par exemple, imaginons que l'on s'intéresse au comportement d'une règle lorsque n_{ab} varie. Si n_{ab} et n_a font partie des paramètres de modélisation choisis, alors on aura tendance à fixer n_a et donc à considérer que $n_{a\bar{b}} = n_a - n_{ab}$ diminue quand n_{ab} augmente. En revanche, si n_{ab} et $n_{a\bar{b}}$ font partie des paramètres choisis, on aura tendance à fixer $n_{a\bar{b}}$ et donc à considérer que $n_a = n_{ab} + n_{a\bar{b}}$ augmente avec n_{ab}.

1.3 Indices de règle

1.3.1 Définition

De nombreux indices de règle ont été développés pour compléter le support et la confiance qui, utilisés seuls, ne permettent d'évaluer que certains aspects de la qualité des règles [Fle96] [Gui00] [BMS97]. Ci-dessous, nous proposons une définition de la notion d'indice de règle.

Définition 1.2 Un **indice de règle** (mesure objective de qualité de règles) est une fonction $I \left| \begin{array}{l} \mathcal{R} \longrightarrow \mathbb{R} \\ (n_{ab}, n_a, n_b, n) \mapsto I(n_{ab}, n_a, n_b, n) \end{array} \right.$ qui est croissante avec n_{ab} et décroissante avec n_a lorsque les autres variables sont fixes. Les variations ne sont pas strictes.

Dans ce chapitre, nous avons recensé un maximum de mesures qui sont traditionnellement utilisées comme indices de règle. Elles sont listées dans le tableau 1.3. Les mesures issues de la théorie de l'information n'y sont pas présentes car elles font l'objet d'une étude spécifique au chapitre 2 (ce sont généralement des mesures de liaison entre variables multimodales et non binaires).

La définition 1.2 est suffisamment générale pour englober toutes les mesures du tableau 1.3. Elle est également suffisamment spécifique pour rejeter les mesures de liaison entre variables multimodales (les symétries de ces mesures font qu'elles ne peuvent satisfaire aux sens de variation de la définition). Un indice de règle doit en effet être une mesure de liaison entre variables binaires puisqu'une règle est une liaison entre variables binaires. Jaroszewicz et Simovici [JS01] soulignent ainsi qu'une règle doit être évaluée uniquement sur les items qui y apparaissent et non sur la distribution jointe complète de la prémisse et de

[4] Il est à noter que les similarités sont généralement modélisées différemment. Dans [Ler81], Lerman choisit une modélisation par les paramètres n_{ab}, $n_{a\bar{b}}$, $n_{\bar{a}b}$, n.

Nom de l'indice *I*	***I(a→b)=***	**Références**
confiance	$\frac{n_{ab}}{n_a}$	[AIS93]
estimateur laplacien de probabilité conditionnelle	$\frac{n_{ab}+1}{n_a+2}$	[BA99]
indice de Sebag et Schoenauer	$\frac{n_{ab}}{n_{a\overline{b}}}$	[SS88]
taux des exemples et contre-exemples	$\frac{n_{ab}-n_{a\overline{b}}}{n_{ab}}$	[Gui04]
indice de Ganascia	$\frac{n_{ab}-n_{a\overline{b}}}{n_a}$	[Gan91]
moindre-contradiction	$\frac{n_{ab}-n_{a\overline{b}}}{n_b}$	[Aze03]
indice d'inclusion	$\sqrt[4]{I^2_{b/a=1} I^2_{\overline{a}/b=0}}$	[GCB+04]
indice de Loevinger	$1 - \frac{n n_{a\overline{b}}}{n_a n_{\overline{b}}}$	[Loe47]
coefficient de corrélation	$\frac{n n_{ab}-n_a n_b}{\sqrt{n_a n_b n_{\overline{a}} n_{\overline{b}}}}$	[Pea96]
rule-interest	$n_{ab} - \frac{n_a n_b}{n}$	[PS91]
nouveauté	$\frac{n_{ab}}{n} - \frac{n_a n_b}{n^2}$	[LFZ99]
lift ou intérêt	$\frac{n n_{ab}}{n_a n_b}$	[BMS97]
conviction	$\frac{n_a n_{\overline{b}}}{n n_{a\overline{b}}}$	[BMUT97]
collective strength	$\frac{n_{ab}+n_{\overline{a}\overline{b}}}{n_a n_b + n_{\overline{a}} n_{\overline{b}}} \frac{n^2 - n_a n_b - n_{\overline{a}} n_{\overline{b}}}{n - n_{ab} - n_{\overline{a}\overline{b}}}$	[AY01]
indice de Yule	$\frac{n_{ab} n_{\overline{a}\overline{b}} - n_{a\overline{b}} n_{\overline{a}b}}{n_{ab} n_{\overline{a}\overline{b}} + n_{a\overline{b}} n_{\overline{a}b}}$	[Yul00]
rapport de cotes	$\frac{n_{ab} n_{\overline{a}\overline{b}}}{n_{a\overline{b}} n_{\overline{a}b}}$	[Mos68]
multiplicateur de cotes	$\frac{n_{ab} n_{\overline{b}}}{n_{a\overline{b}} n_b}$	[LT04]
κ	$\frac{n n_{ab} + n n_{\overline{a}\overline{b}} - n_a n_b - n_{\overline{a}} n_{\overline{b}}}{n^2 - n_a n_b - n_{\overline{a}} n_{\overline{b}}}$	[Coh60]
intensité d'implication	$\mathrm{P}(Poisson(\frac{n_a n_{\overline{b}}}{n}) > n_{a\overline{b}})$	[Gra96]
indice de vraisemblance du lien	$\mathrm{P}(Poisson(\frac{n_a n_b}{n}) < n_{ab})$	[Ler81]
opposé de l'indice d'implication	$-\frac{n_{a\overline{b}} - \frac{n_a n_{\overline{b}}}{n}}{\sqrt{\frac{n_a n_{\overline{b}}}{n}}}$	[Gra96]
contribution orientée au χ^2	$\frac{n_{ab} - \frac{n_a n_b}{n}}{\sqrt{\frac{n_a n_b}{n}}}$	[Ler81]
support ou indice de Russel et Rao	$\frac{n_{ab}}{n}$	[AIS93] [RR40]
support causal ou indice de Sokal et Michener	$\frac{n_{ab}+n_{\overline{a}\overline{b}}}{n}$	[Kod00] [SM58]
indice de Rogers et Tanimoto	$\frac{n - n_{a\overline{b}} - n_{\overline{a}b}}{n + n_{a\overline{b}} + n_{\overline{a}b}}$	[RT60]
indice de Jaccard	$\frac{n_{ab}}{n - n_{\overline{a}\overline{b}}}$	[Jac01]
indice de Dice	$\frac{n_{ab}}{n_{ab} + \frac{1}{2}(n_{a\overline{b}} + n_{\overline{a}b})}$	[Dic45]
indice d'Ochiai	$\frac{n_{ab}}{\sqrt{n_a n_b}}$	[Och57]
indice de Kulczynski	$\frac{1}{2}(\frac{n_{ab}}{n_a} + \frac{n_{ab}}{n_b})$	[Kul27]

Les fonctions entropiques $I_{b/a=1}$ et $I_{\overline{a}/b=0}$ sont définies chapitre 2 page 50.

TAB. 1.3 – Les principaux indices de règle

la conclusion, comme le font les mesures de liaison entre variables multimodales. Appliquée à une règle, une mesure de liaison entre variables multimodales ne différencierait pas les exemples des contre-exemples, puisqu'elle traite identiquement vérité et fausseté (voir section 1.2.1). Par là même, elle ne permettrait pas non plus de distinguer les règles contraires $a \rightarrow b$ et $a \rightarrow \overline{b}$ alors même qu'elles ont des significations opposées.

Les sens de variation d'un indice de règle avec n_{ab} et n_a ont été soulignés à l'origine par Piatetsky-Shapiro [PS91] en tant que propriétés souhaitables d'un indice. Nous les considérons ici comme les fondements de la notion d'indice de règle. Piatetsky-Shapiro considère aussi qu'un bon indice doit décroître avec n_b. Cette condition est trop contraignante pour apparaître dans une définition générale des indices de règle comme la définition 1.2, puisque certains indices ne dépendent pas de n_b. Plus généralement, en ce qui concerne les variations par rapport à n_b et n, un indice de règle n'a pas de comportement précis :

- Certains indices ne dépendent pas de n, comme le taux d'exemples et de contre-exemples, la moindre-contradiction, ou l'indice d'Ochiai.
- Certains indices croissent avec n, comme l'indice d'inclusion, l'indice de Yule, l'indice de Rogers et Tanimoto, rule-interest.
- Certains indices décroissent avec n comme le support.
- Certains indices ne sont pas monotones avec n comme la nouveauté.
- Certains indices ne dépendent pas de n_b, comme la confiance, l'indice de Sebag et Schoenauer, le support.
- Certains indices décroissent avec n_b, comme le multiplicateur de cotes, collective strength, l'indice de Jaccard (nous verrons par la suite que tous les indices que nous qualifions d'*écart à l'indépendance* décroissent avec n_b).

1.3.2 Comparaison aux indices de similarité

Il existe en analyse de données une famille de mesures nommées indices de similarité, utilisée pour l'étude d'individus décrits par des variables binaires (appelées caractères). Les indices de similarité ont pour but d'évaluer la ressemblance entre deux individus ou deux caractères. Lerman en donne la définition suivante [Ler81].

Définition 1.3 Nous notons $\mathcal{S}$ le sous-ensemble de $\mathbb{N}^4$ suivant : $\mathcal{S} = \left\{ (n_{ab},\ n_{a\overline{b}},\ n_{\overline{a}b},\ n) \mid n_{ab} + n_{a\overline{b}} + n_{\overline{a}b} \leq n \right\}$. Un **indice de similarité** est une fonction $I_s \left| \begin{array}{l} \mathcal{S} \longrightarrow \mathbb{R} \\ (n_{ab}, n_{a\overline{b}}, n_{\overline{a}b}, n) \mapsto I_s(n_{ab}, n_{a\overline{b}}, n_{\overline{a}b}, n) \end{array} \right.$ qui est positive, symétrique en $n_{a\overline{b}}$ et $n_{\overline{a}b}$, croissante avec n_{ab} et décroissante avec $n_{a\overline{b}}$ lorsque les autres variables sont fixes. Les variations sont strictes.

Dans le tableau 1.3, les indices de Russel et Rao (support), Sokal et Michener (support causal), Rogers et Tanimoto, Jaccard, Dice, Ochiai, et Kulczynski sont des indices de similarité. Ci-dessous, nous montrons qu'un indice de similarité est un indice de règle.

Démonstration 1.1 *Soit I_s un indice de similarité. Etant donné $(n_{ab},\ n_a,\ n_b,\ n) \in \mathcal{R}$, nous avons $(n_{ab},\ n_a - n_{ab},\ n_b - n_{ab},\ n) \in \mathcal{S}$. Nous pouvons donc définir la fonction I suivante de $\mathcal{R}$ dans $\mathbb{R}$:*

$$\forall\ (n_{ab}, n_a, n_b, n) \in \mathcal{R},\ I(n_{ab}, n_a, n_b, n) = I_s(n_{ab}, n_a - n_{ab}, n_b - n_{ab}, n)$$

La fonction I est un indice de règle si elle croît avec n_{ab} et décroît avec n_a lorsque les autres variables sont fixes. Faisons croître n_{ab} avec n_a, n_b, et n constants. $n_a - n_{ab}$ et $n_b - n_{ab}$ décroissent. Or I_s croît avec sa première variable et décroît avec ses deuxième et troisième variables. Donc I croît.
Faisons croître n_a avec n_{ab}, n_b, et n constants. I_s décroît avec sa deuxième variable, donc I décroît.

En revanche, un indice de règle, même positif et symétrique par permutation des variables, n'est pas un indice de similarité. Par exemple, le lift peut décroître lorsque n_{ab} augmente avec $n_{a\bar{b}}$, $n_{\bar{a}b}$, et n constants.

1.3.3 Evaluation et sélection des indices

Plusieurs auteurs se sont intéressés aux propriétés qu'objectivement un bon indice doit vérifier [PS91] [Fre99] [TKS04] [LT04] [GCB+04]. Dans le cadre de leur participation au groupe de travail GafoQualité de l'action spécifique STIC [BSGG04], Gras *et al.* et Lallich et Teytaud proposent plusieurs critères d'évaluation des indices, dont les principaux sont :

- valeur de l'indice quand les variables a et b sont indépendantes,
- valeur de l'indice pour une règle logique,
- variations de l'indice avec n_b,
- variations de l'indice quand tous les effectifs sont dilatés,
- symétrie de l'indice pour une règle et sa réciproque,
- symétrie de l'indice pour une règle et sa règle contraire,
- symétrie de l'indice pour une règle et sa contraposée,
- concavité/convexité de l'indice pour $n_{a\bar{b}} = 0^+$,
- intelligibilité de l'indice,
- facilité à fixer un seuil d'acceptation des règles,
- sensibilité de l'indice aux règles très spécifiques.

La classification des indices de règle que nous proposons à la section 1.4 réutilise certains des critères ci-dessus (ceux qui nous paraissent essentiels pour saisir la signification des indices) tout en en adjoignant d'autres.

D'autres travaux présentent des études comparatives formelles ou expérimentales[5] des indices [BA99] [TK00] [TKS04] [VLL04]. Bayardo [BA99] montre que pour n_b fixé, un certain nombre d'indices sont redondants et il suffit d'utiliser le support et la confiance pour mettre en évidence les meilleures règles. Tan *et al.* [TK00] [TKS04] comparent entre eux des indices symétriques ou symétrisées[6]

[5] Les approches expérimentales consistent à comparer les indices sur des jeux de règles. Les résultats dépendent cependant des données dont sont issues les règles et des biais induits par les paramètres des algorithmes d'extraction de règles (seuils de support et de confiance, nombre maximal d'items dans une règle, prise en compte ou non des négations d'items).

[6] Les indices n'évaluent pas une règle au sens strict puisqu'ils jugent de la même façon une règle et sa réciproque : $I(a \rightarrow b) = I(b \rightarrow a)$.

selon différents critères formels et sur des jeux de règles synthétiques. Les indices se révèlent tantôt redondants, tantôt contradictoires, et aucun ne surclasse significativement tous les autres. Vaillant *et al.* [VLL04] quant à eux réalisent une étude expérimentale approfondie en comparant une vingtaine d'indices sur différentes bases de règles. Ils mettent en évidence trois groupes d'indices similaires globalement stables sur les bases considérées.

Tous ces travaux s'accordent à dire qu'il n'existe pas d'indice idéal. Certaines caractéristiques des indices peuvent être appropriées à certaines applications sur certaines données pour certains utilisateurs, mais pas à d'autres applications sur d'autres données pour d'autres utilisateurs. Plus précisément, la pertinence d'un indice dépend de l'idée que l'utilisateur se fait d'une règle de bonne qualité sur les données étudiées et pour son application. Le choix des indices à appliquer sur les règles doit donc être effectué par l'utilisateur, sinon avec lui, en fonction de ses préférences (l'utilisation d'indices objectifs de qualité n'annule pas le caractère hautement subjectif du post-traitement de règles). Afin de l'assister dans sa sélection parmi la multitude d'indices candidats, Lenca *et al.* [LMV+04] proposent d'appliquer une méthode d'aide multicritère à la décision. Après que l'utilisateur ait exprimé ses préférences sur des critères tels que ceux listés plus haut, la méthode lui fait une recommandation sous la forme d'un ordre partiel sur les indices.

1.4 Classification des indices de règle

Dans cette section, nous proposons une classification des indices de règle selon trois critères : l'objet de l'indice, la portée de l'indice, et la nature de l'indice. Ces critères nous paraissent essentiels pour appréhender la signification des indices, et donc aussi pour aider l'utilisateur à choisir quels indices appliquer.

1.4.1 Objet d'un indice de règle

Nous avons vu qu'une règle est d'autant meilleure qu'elle admet beaucoup d'exemples et peu de contre-exemples. Ainsi, pour n_a, n_b et n donnés, la qualité de $a \rightarrow b$ est maximale lorsque $n_{ab} = \min(n_a,\ n_b)$ et minimale lorsque $n_{ab} = \max(0,\ n_a + n_b - n)$. Entre ces situations extrêmes, il existe deux configurations intéressantes dans lesquelles les règles apparaissent comme des liaisons non orientées et peuvent donc être considérées comme neutres ou inexistantes : l'indépendance et l'équilibre. Une règle qui se trouve dans l'une de ces configurations est à rejeter.

Indépendance

Les variables binaires a et b sont indépendantes si et seulement si $\mathrm{P}(a \cap b) = \mathrm{P}(a) \times \mathrm{P}(b)$, soit $\mathbf{n.n_{ab}} = \mathbf{n_a n_b}$. Dans ce cas, chaque variable n'apporte aucune information sur l'autre, puisque la connaissance de la modalité prise par l'une des variables n'influe pas sur la distribution de probabilités de l'autre variable : $\mathrm{P}(b \backslash a) = \mathrm{P}(b \backslash \overline{a}) = \mathrm{P}(b)$ et $\mathrm{P}(\overline{b} \backslash a) = \mathrm{P}(\overline{b} \backslash \overline{a}) = \mathrm{P}(\overline{b})$ (idem pour les probabilités de a et $\overline{a}$ sachant b ou $\overline{b}$). En d'autres termes, la connaissance de la modalité prise

a \ b	**1**	**0**	
1	$\frac{n_a \times n_b}{n}$	$\frac{n_a \times n_{\overline{b}}}{n}$	n_a
0	$\frac{n_{\overline{a}} \times n_b}{n}$	$\frac{n_{\overline{a}} \times n_{\overline{b}}}{n}$	$n_{\overline{a}}$
	n_b	$n_{\overline{b}}$	n

TAB. 1.4 – Effectifs à l'indépendance entre a et b

a \ b	**1**	**0**		
1	$\frac{n_a \times n_b}{n} + \Delta$	$\frac{n_a \times n_{\overline{b}}}{n} - \Delta$	n_a	$(\Delta > 0)$
0	$\frac{n_{\overline{a}} \times n_b}{n} - \Delta$	$\frac{n_{\overline{a}} \times n_{\overline{b}}}{n} + \Delta$	$n_{\overline{a}}$	
	n_b	$n_{\overline{b}}$	n	

TAB. 1.5 – Corrélation positive entre a et b

a \ b	**1**	**0**		
1	$\frac{n_a \times n_b}{n} - \Delta$	$\frac{n_a \times n_{\overline{b}}}{n} + \Delta$	n_a	$(\Delta > 0)$
0	$\frac{n_{\overline{a}} \times n_b}{n} + \Delta$	$\frac{n_{\overline{a}} \times n_{\overline{b}}}{n} - \Delta$	$n_{\overline{a}}$	
	n_b	$n_{\overline{b}}$	n	

TAB. 1.6 – Corrélation négative entre a et b

par l'une des variables laisse intacte notre incertitude concernant la modalité prise par l'autre variable.

Pour deux variables a et b données, il existe une unique situation d'indépendance, commune aux huit règles $a \rightarrow b$, $a \rightarrow \overline{b}$, $\overline{a} \rightarrow b$, $\overline{a} \rightarrow \overline{b}$, $b \rightarrow a$, $b \rightarrow \overline{a}$, $\overline{b} \rightarrow a$, et $\overline{b} \rightarrow \overline{a}$. La table de contingence de deux variables indépendantes est donnée tableau 1.4. Il existe deux façons de s'écarter de la situation d'indépendance :

- soit les variables a et b sont corrélées positivement ($\mathrm{P}(a \cap b) > \mathrm{P}(a) \times \mathrm{P}(b)$), et ce sont alors les quatre règles $a \rightarrow b$, $\overline{a} \rightarrow \overline{b}$, $b \rightarrow a$, et $\overline{b} \rightarrow \overline{a}$ qui apparaissent dans les données ;
- soit les variables a et b sont corrélées négativement ($\mathrm{P}(a \cap b) < \mathrm{P}(a) \times \mathrm{P}(b)$), et ce sont alors les quatre règles contraires qui apparaissent dans les données : $a \rightarrow \overline{b}$, $\overline{a} \rightarrow b$, $b \rightarrow \overline{a}$, et $\overline{b} \rightarrow a$.

Cette dichotomie entre les règles et leurs règles contraires s'explique par le fait que deux règles contraires sont contravariantes, puisque les exemples de l'une sont les contre-exemples de l'autre, et réciproquement. Ainsi, si l'une possède plus d'exemples et donc moins de contre-exemples qu'à l'indépendance, alors l'autre possède moins d'exemples et plus de contre-exemples qu'à l'indépendance, et réciproquement. Les tables de contingence 1.5 et 1.6 illustrent respectivement une corrélation positive et une corrélation négative entre a et b avec n_a, n_b et n fixés et un degré de liberté unique noté Δ. Les règles sont représentées sur les tables par des flèches. Par exemple, une flèche de la cellule $n_{a\overline{b}}$ vers la cellule n_{ab} illustre la migration des effectifs (par rapport à la situation d'indépendance) de la cellule $n_{a\overline{b}}$ vers la cellule n_{ab}, et représente donc la règle $a \rightarrow b$.

Equilibre

Nous définissons l'équilibre d'une règle $a \rightarrow b$ comme la situation où la règle possède autant d'exemples que de contre-exemples : $\mathbf{n_{ab} = n_{a\overline{b}} = \frac{1}{2} n_a}$ [BGGB04] [BGBG05a]. Dans cette situation, l'événement $a = 1$ est dans les données autant concomitant avec $b = 1$ qu'avec $b = 0$. Une règle $a \rightarrow b$ à l'équilibre est donc autant orientée vers b que vers $\overline{b}$.

a \ b	**1**	**0**	
1	$\frac{n_a}{2}$	$\frac{n_a}{2}$	n_a
	n_{ab}	$n_{a\overline{b}}$	n_a

TAB. 1.7 – Effectifs à l'équilibre de b vis-à-vis de $a = 1$

L'équilibre $n_{ab} = n_{a\overline{b}}$ est une situation définie non pas pour les deux variables a et b mais pour la variable b vis-à-vis du littéral $a = 1$. Ainsi avec deux variables, il existe quatre équilibres différents, chacun étant commun à deux règles contraires. Le tableau 1.7 donne les effectifs pour l'équilibre de b vis-à-vis de $a = 1$ sous la forme d'une demi-table de contingence, tandis que les tableaux 1.8 et 1.9 montrent les deux façons de s'écarter de cette situation d'équilibre

a \ b	**1**	**0**	
1	$\frac{n_a}{2} + \Delta$	$\frac{n_a}{2} - \Delta$	n_a
	n_{ab}	$n_{a\bar{b}}$	n_a

$(\Delta > 0)$

TAB. 1.8 – Déséquilibre de b en faveur de $b = 1$ vis-à-vis de $a = 1$

a \ b	**1**	**0**	
1	$\frac{n_a}{2} - \Delta$	$\frac{n_a}{2} + \Delta$	n_a
	n_{ab}	$n_{a\bar{b}}$	n_a

$(\Delta > 0)$

TAB. 1.9 – Déséquilibre de b en faveur de $b = 0$ vis-à-vis de $a = 1$

avec n_b fixé et un degré de liberté noté Δ. Si $a = 1$ est plus concomitant avec $b = 1$ qu'avec $b = 0$ alors c'est la règle $a \rightarrow b$ qui apparaît dans les données, sinon il s'agit de $a \rightarrow \overline{b}$. On retrouve ici la contravariance des règles contraires : si l'une est meilleure qu'à l'équilibre alors l'autre est moins bonne qu'à l'équilibre, et réciproquement.

Indices d'écart à l'indépendance et d'écart à l'équilibre

L'existence de deux notions de neutralité différentes pour les règles montre que la qualité objective des règles doit être évaluée selon (au moins) deux points de vue complémentaires : l'écart à l'indépendance et l'écart à l'équilibre. Il s'agit d'écarts **orientés**, en faveur des exemples et en défaveur des contre-exemples. Du point de vue de l'écart à l'équilibre, une règle $a \rightarrow b$ avec un bon écart signifie :

"Quand a est vrai, alors b est très souvent vrai."

Du point de vue de l'écart à l'indépendance en revanche, une règle $a \rightarrow b$ avec un bon écart signifie :

"Quand a est vrai, alors b est plus souvent vrai (qu'à l'accoutumée)."

L'écart à l'équilibre est un constat absolu, alors que l'écart à l'indépendance est une comparaison relativement à une situation attendue (caractérisée par n_b).

Définition 1.4 Un indice de règle I mesure un **écart à l'indépendance** si et seulement si l'indice prend une valeur fixe (indépendante des données) à l'indépendance :

$$I(\frac{n_a n_b}{n}, n_a, n_b, n) = constante$$

Définition 1.5 Un indice de règle I mesure un **écart à l'équilibre** si et seulement si l'indice prend une valeur fixe à l'équilibre :

$$I(\frac{n_a}{2}, n_a, n_b, n) = constante$$

D'un point de vue général, un indice d'écart à l'équilibre est utile pour prendre des décisions sur b ou faire des prédictions sur b (sachant ou imaginant que a est vrai, b est-il vrai ou faux ?), tandis qu'un indice d'écart à l'indépendance est utile pour découvrir des liaisons entre a et b (la vérité de a influence-t-elle la vérité de b ?).

Exemple. Considérons une règle $fumer \rightarrow cancer$ (fumer provoque le cancer) évaluée au moyen de la confiance et du lift. Son écart à l'équilibre est mesuré par une confiance de 40%, ce qui signifie que 40% des personnes qui fument développent un cancer ; son écart à l'indépendance est mesuré par un lift de 10, ce qui signifie que fumer augmente les risques de développer un cancer d'un facteur 10. Une personne qui fume et qui cherche à savoir si elle risque de développer un cancer est plutôt intéressée par l'écart à l'équilibre. En revanche, une personne qui ne fume pas mais qui hésite à commencer est plutôt intéressée par l'écart à l'indépendance. □

L'indépendance se définit à l'aide des quatre paramètres n_{ab}, n_a, n_b, et n, alors que l'équilibre ne se définit qu'à l'aide des paramètres n_{ab} et n_a. Ainsi, les indices d'écart à l'indépendance sont des fonctions des quatre paramètres (en particulier, ils décroissent tous avec n_b). En revanche, les indices d'écart à l'équilibre ne sont généralement pas fonctions de n_b et de n. Les seules exceptions à ce principe sont l'indice d'inclusion et la moindre-contradiction :

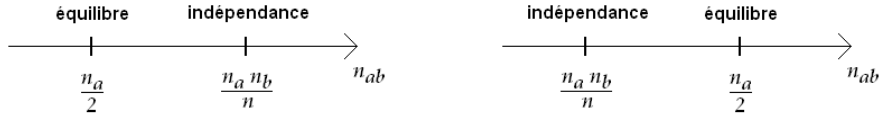
- L'indice d'inclusion est fonction de n_b et n. Il associe en effet deux indices de règle, l'un relatif à la règle directe (fonction de n_{ab} et n_a), et l'autre à la règle contraposée (fonction de $n_{\overline{a}\overline{b}}$ et $n_{\overline{b}}$, c'est-à-dire de $n - n_a - n_b + n_{ab}$ et de $n - n_b$). C'est la contribution de la contraposée qui introduit une dépendance envers n_b et n dans l'indice d'inclusion. Cet indice est présenté en détails au chapitre 2.
- La moindre-contradiction est fonction de n_b. C'est un indice hybride qui, par son numérateur, s'annule à l'équilibre comme les indices d'écart à l'équilibre, et, par son dénominateur, décroît avec n_b comme les indices d'écart à l'indépendance.

Dans la pratique, les indices d'écart à l'indépendance et à l'équilibre facilitent la validation des règles. Tout d'abord, ils permettent de repérer aisément les règles qui doivent nécessairement être rejetées :

- pour un indice d'écart à l'indépendance I_{idp} prenant la valeur v_{idp} à l'indépendance, si une règle r vérifie $I_{idp}(r) \leq v_{idp}$, alors r doit être rejetée car elle est établie entre deux variables corrélées négativement ;
- pour un indice d'écart à l'équilibre I_{eql} prenant la valeur v_{eql} à l'équilibre, si une règle r vérifie $I_{eql}(r) \leq v_{eql}$, alors r doit être rejetée car elle possède plus de contre-exemples que d'exemples.

Ensuite, les indices d'écart à l'indépendance et à l'équilibre facilitent la fixation d'un seuil σ pour le filtrage des règles. En effet, pour que ce seuil soit pertinent, il doit vérifier $\sigma > v_{idp}$ si le filtrage porte sur les valeurs prises par un indice d'écart à l'indépendance, et $\sigma > v_{eql}$ si le filtrage porte sur les valeurs prises par un indice d'écart à l'équilibre.

Les notions d'écarts à l'équilibre et à l'indépendance ne sont pas nouvelles en tant que telles. D'une part, de nombreux indices de règle mesurent un écart à



(A) $n_b \geq \frac{n}{2}$ (B) $n_b \leq \frac{n}{2}$

FIG. 1.4 – Deux cas de figure possibles pour l'équilibre et l'indépendance

l'équilibre, aux premiers desquels la confiance. Dans la littérature sur les règles d'association, l'écart à l'équilibre est souvent appelé "force", "puissance", "validité", ou "qualité inclusive" de la règle. D'autres part, les indices qui permettent de mesurer un écart à l'indépendance sont encore plus nombreux. La mesure de l'écart à l'indépendance est d'ailleurs classique en statistique.

Comparaison des préordonnances

Soient I_{idp} et I_{eql} deux indices de règle qui mesurent respectivement un écart à l'indépendance et un écart à l'équilibre. Les valeurs fixes prises à l'indépendance et à l'équilibre sont notées respectivement v_{idp} et v_{eql} :

$$I_{idp}(\frac{n_a n_b}{n}, n_a, n_b, n) = v_{idp} \quad (1.1)$$

$$I_{eql}(\frac{n_a}{2}, n_a, n_b, n) = v_{eql} \quad (1.2)$$

Nous cherchons ici à exhiber deux règles r_1 et r_2 qui sont ordonnées différemment par I_{idp} et I_{eql}, c'est-à-dire $I_{idp}(r_1) \leq I_{idp}(r_2)$ et $I_{eql}(r_1) \geq I_{eql}(r_2)$. Pour ce faire, nous mettons ci-dessous en évidence deux familles de règles qui sont toujours ordonnées différemment.

Considérons une règle $(n_{ab},\ n_a,\ n_b,\ n)$. En faisant varier n_{ab} avec n_a, n_b, et n fixes, on peut distinguer deux cas de figure différents pour la règle (voir figure 1.4) [BGGB04] [BGBG05a] :

- Si $n_b \geq \frac{n}{2}$ (cas 1), alors $\frac{n_a n_b}{n} \geq \frac{n_a}{2}$, et donc la règle passe à l'équilibre avant de passer à l'indépendance quand n_{ab} augmente.
- Si $n_b \leq \frac{n}{2}$ (cas 2), alors $\frac{n_a n_b}{n} \leq \frac{n_a}{2}$, et donc la règle passe à l'indépendance avant de passer à l'équilibre quand n_{ab} augmente.

Supposons que n_{ab} soit compris entre $\frac{n_a}{2}$ et $\frac{n_a n_b}{n}$. La règle est entre équilibre et indépendance. Plus précisément :

- Dans le cas 1, on a $\frac{n_a}{2} \leq n_{ab} \leq \frac{n_a n_b}{n}$. Un indice de règle étant une fonction croissante avec n_{ab} lorsque les autres variables sont fixes, on obtient d'après 1.1 et 1.2 :

  $$I_{idp}(n_{ab}, n_a, n_b, n) \leq v_{idp} \quad et \quad I_{eql}(n_{ab}, n_a, n_b, n) \geq v_{eql} \quad (1.3)$$

  La règle est à rejeter du point de vue de son écart à l'indépendance mais elle est acceptable du point de vue de son écart à l'équilibre.

– Dans le cas 2, on a $\frac{n_a n_b}{n} \leq n_{ab} \leq \frac{n_a}{2}$. De la même façon, on obtient d'après 1.1 et 1.2 :

$$I_{idp}(n_{ab}, n_a, n_b, n) \geq v_{idp} \quad et \quad I_{eql}(n_{ab}, n_a, n_b, n) \leq v_{eql} \tag{1.4}$$

La règle est à rejeter du point de vue de son écart à l'équilibre mais elle est acceptable du point de vue de son écart à l'indépendance.

Ainsi, pour exhiber deux règles r_1 et r_2 qui sont ordonnées différemment par I_{idp} et I_{eql}, il suffit de choisir r_1 entre équilibre et indépendance dans le cas 1 et r_2 entre équilibre et indépendance dans le cas 2, c'est-à-dire :

$$r_1 = (n_{ab_1},\ n_{a_1},\ n_{b_1},\ n_1) \ \ avec \ \ \frac{n_{a_1}}{2} \leq n_{ab_1} \leq \frac{n_{a_1} n_{b_1}}{n_1} \ \ et \ \ n_{b_1} \geq \frac{n_1}{2}$$

$$r_2 = (n_{ab_2},\ n_{a_2},\ n_{b_2},\ n_2) \ \ avec \ \ \frac{n_{a_2} n_{b_2}}{n_2} \leq n_{ab_2} \leq \frac{n_{a_2}}{2} \ \ et \ \ n_{b_2} \leq \frac{n_2}{2}$$

($n_1 = n_2$ si l'on désire choisir deux règles issues du même jeu de données)

Les inégalités 1.3 et 1.4 appliquées respectivement à r_1 et r_2 donnent :

$$I_{idp}(r_1) \leq v_{idp} \leq I_{idp}(r_2) \quad et \quad I_{eql}(r_2) \leq v_{eql} \leq I_{eql}(r_1)$$

Exemple. Considérons les règles r_1 = (800, 1000, 4500, 5000) et r_2 = (400, 1000, 1000, 5000) évaluées par la confiance (indice d'écart à l'équilibre) et le lift (indice d'écart à l'indépendance). La confiance vaut 0.5 à l'équilibre, et le lift vaut 1 à l'indépendance. Il vient :

$$confiance(r_1) = 0.8 \quad \text{et} \quad lift(r_1) = 0.9$$

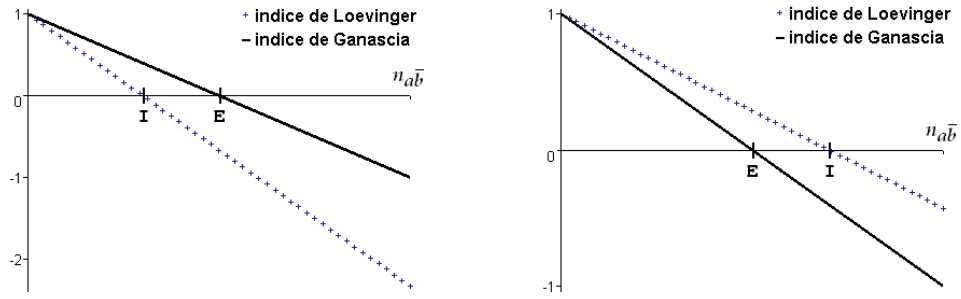
$$confiance(r_2) = 0.4 \quad \text{et} \quad lift(r_2) = 2$$

La règle r_1 est bien évaluée par la confiance mais mal évaluée par le lift, alors que la règle r_2 est bien évaluée par le lift mais mal évaluée par la confiance. On a $confiance(r_1) \geq confiance(r_2)$ et $lift(r_1) \leq lift(r_2)$. □

Un indice d'écart à l'indépendance et un indice d'écart à l'équilibre n'induisent donc pas le même préordre sur $\mathcal{R}$. Ceci montre qu'un indice de règle (à moins d'être constant) ne peut pas mesurer à la fois un écart à l'indépendance et un écart à l'équilibre. Cela confirme également que les deux écarts sont deux aspects différents de la qualité des règles : l'écart à l'équilibre peut être bon sans que l'écart à l'indépendance ne le soit, et réciproquement. Etonnamment, même si cette idée est sous-jacente à divers travaux concernant les règles d'association, il nous semble qu'il n'a jamais été clairement énoncé que la qualité objective des règles réside de façon complémentaire dans les deux notions d'écart à l'indépendance et d'écart à l'équilibre.

Comparaison des filtres

Nous comparons maintenant les pouvoirs filtrant d'un indice d'écart à l'équilibre I_{eql} et un indice d'écart à l'indépendance I_{idp} (utiliser un indice de règle



(A) cas 1 ($n_b \geq \frac{n}{2}$) (B) cas 2 ($n_b \leq \frac{n}{2}$)

FIG. 1.5 – Comparaison des indices de Ganascia et Loevinger
(E : équilibre, I : indépendance)

en tant que filtre consiste à ne retenir que les règles qui respectent un seuil minimal). Afin que la comparaison soit équitable, nous supposons que les deux indices ont des comportements similaires :

- même valeur pour une règle logique,
- même valeur à l'équilibre/indépendance,
- même vitesse de décroissance en fonction des contre-exemples.

Par exemple, I_{eql} peut être l'indice de Ganascia et I_{idp} l'indice de Loevinger. Les cas de figure 1 et 2 présentés à la section précédente sont possibles :

- Dans le cas 1, on a $\frac{n_a n_{\bar{b}}}{n} \leq \frac{n_a}{2}$. Cela signifie que pour n_a, n_b et n donnés, la quantité de contre-exemples nécessaire pour atteindre l'indépendance ($n_{a\bar{b}} = \frac{n_a n_{\bar{b}}}{n}$) est plus faible que la quantité de contre-exemples nécessaire pour atteindre l'équilibre ($n_{a\bar{b}} = \frac{n_a}{2}$). Dans le cas 1, I_{idp} est donc plus filtrant que I_{eql}, comme l'illustre la figure 1.5.(A).
- Dans le cas 2, on a $\frac{n_a n_{\bar{b}}}{n} \geq \frac{n_a}{2}$. Cela signifie que pour n_a, n_b et n donnés, la quantité de contre-exemples nécessaire pour atteindre l'équilibre est plus faible que la quantité de contre-exemples nécessaire pour atteindre l'indépendance. Dans le cas 2, I_{eql} est donc plus filtrant que I_{idp}, comme l'illustre la figure 1.5.(B).

En d'autres termes, dans le cas 1 c'est I_{idp} qui contribue à rejeter les mauvaises règles, tandis que dans le cas 2 c'est I_{eql}. Ceci montre que les indices d'écart à l'équilibre et d'écart à l'indépendance doivent être considérés comme complémentaires. En particulier, il ne faut pas négliger les indices d'écart à l'équilibre au profit des indices d'écart à l'indépendance quand les réalisations des variables étudiées sont rares. Dans cette situation, en effet, pour peu que l'utilisateur ne s'intéresse pas aux règles portant sur les non-réalisations (ce qui en général se vérifie dans la pratique), le cas 2 est plus fréquent que le cas 1. De nombreux auteurs citent pourtant parmi les propriétés majeures d'un bon indice de règle le principe suivant (énoncé à l'origine dans [PS91]) : "un indice doit s'annuler (ou prendre une valeur fixe) à l'indépendance" [TKS04] [GCB+04] [LT04] [LMV+04]. Ce principe nie totalement la notion d'écart à l'équilibre ; l'utiliser revient à considérer que les indices d'écart à l'indépendance mesurent mieux la qualité des règles que les indices d'écart à l'équilibre.

Indices d'écart à l'équilibre	**Indices d'écart à l'indépendance**	**Indices de similarité**
confiance	indice de Loevinger	support ou indice de Russel et Rao
indice de Sebag et Schoenauer	coefficient de corrélation	support causal ou indice de Sokal et Michener
taux des exemples et contre-exemples	rule-interest	indice de Rogers et Tanimoto
estimateur laplacien de probabilité conditionnelle	nouveauté	indice de Jaccard
indice de Ganascia	lift ou intérêt	indice de Dice
moindre-contradiction	conviction	indice d'Ochiai
indice d'inclusion	collective strength	indice de Kulczynski
	indice de Yule	
	rapport de cotes	
	multiplicateur de cotes	
	κ	
	intensité d'implication	
	indice de vraisemblance du lien	
	opposé de l'indice d'implication	
	contribution orientée au χ^2	

TAB. 1.10 – Classification des indices de règle selon leur objet

Classification des indices selon l'objet

Dans notre classification, l'objet d'un indice est la notion mesurée par celui-ci. Il peut s'agir d'un écart à l'équilibre ou d'un écart à l'indépendance (voir tableau 1.10). Il existe cependant des indices qui ne mesurent aucun des deux écarts : les indices de similarité[7] de Rogers et Tanimoto, Jaccard, Dice, Ochiai, Kulczynski, Russel et Rao (support), et Sokal et Michener (support causal). Dans la classification, nous créons donc une troisième catégorie pour les indices de similarité. Parmi eux, le support met en évidence un aspect supplémentaire de la qualité des règles : leur généralité/spécificité. Le support causal peut également être considéré comme une mesure de généralité/spécificité, mais pour une règle et sa contraposée. Les indices d'Ochiai et de Kulczynski sont respectivement les moyennes géométriques et arithmétiques entre les confiances de $a \rightarrow b$ et de $b \rightarrow a$. La signification des autres indices de similarité ne peut pas s'exprimer en termes de règle de manière évidente.

1.4.2 Portée d'un indice de règle

Quasi-implication

De prime abord, on peut penser qu'une règle est une approximation de l'implication logique (ou implication matérielle) qui admettrait des contre-exemples, ce qu'on appelle communément une "quasi-implication". Cependant, une règle et une implication ne sont en fait pas si analogues. Il suffit pour s'en persuader de comparer les tableaux 1.11.(A) et (B), qui représentent respectivement la table de contingence de la règle $a \rightarrow b$ et la table de vérité de l'implication logique $a \supset b$. Les cas qui possèdent le même rôle pour la règle et l'implication sont les cas ($a = 1$ et $b = 1$) et ($a = 1$ et $b = 0$) : les premiers vérifient la règle et l'implication, et les seconds les contredisent. En revanche, les cas ($a = 0$ et $b = 1$) et ($a = 0$ et $b = 0$) ne jouent pas le même rôle pour $a \rightarrow b$ et $a \supset b$: ils vérifient l'implication mais ne sont pas pour autant des exemples pour la règle. En fait, pour $a \rightarrow b$, ces cas n'ont pas de rôle défini [LT04]. Une règle traduit en effet uniquement la tendance de la conclusion à être vraie quand la prémisse est vraie. Le fait que les cas ($a = 0$ et $b = 1$) et ($a = 0$ et $b = 0$) vérifient l'implication est souvent considéré comme un paradoxe de l'implication logique (une prémisse fausse peut impliquer n'importe quelle conclusion). Des paradoxes comme celui-ci ont motivé le développement de logiques non classiques visant à représenter plus fidèlement la "logique" de sens commun, et dans lesquelles l'implication ne donne pas lieu (ou donne moins lieu) à des énoncés contre-intuitifs. Il s'agit des logiques modales et des logiques conditionnelles (voir [Gui99] pour une introduction), ainsi que des logiques de pertinence [Dun86].

Une implication logique $a \supset b$ est équivalente à sa contraposée $\overline{b} \supset \overline{a}$. Ainsi, il est possible d'effectuer des déductions :

- soit en affirmant la prémisse a (*Modus ponens*) – c'est la forme directe de l'implication qui est utilisée ($a \supset b$),

[7] Remarquez que certains indices mesurant un écart à l'indépendance sont parfois considérés comme faisant partie des indices de similarité (indice de Yule, corrélation). Nous préférons les classer à part puisque selon nous, ils évaluent plus qu'une simple ressemblance entre caractères.

a \ b	**1**	**0**	
1	exemples n_{ab}	contre-exemples $n_{a\overline{b}}$	n_a
0	$n_{\overline{a}b}$	$n_{\overline{a}\overline{b}}$	$n_{\overline{a}}$
	n_b	$n_{\overline{b}}$	n

(A) Table de contingence de la règle $a \rightarrow b$

a \ b	**1**	**0**
1	1	0
0	1	1

(B) Table de vérité de l'implication logique $a \supset b$

TAB. 1.11 – Comparaison d'une règle avec l'implication logique

– soit en niant la conclusion b (*Modus tollens*) – c'est la forme contraposée de l'implication qui est utilisée.

Par contre, dans le cas général, une règle $a \rightarrow b$ n'est pas équivalente à sa contraposée $\overline{b} \rightarrow \overline{a}$: la tendance de la conclusion à être vraie quand la prémisse est vraie n'est pas forcément identique à la tendance de la prémisse à être fausse quand la conclusion est fausse. En particulier, certains indices de règle peuvent mesurer des qualités très différentes pour une règle et sa contraposée[8]. L'utilisateur peut cependant interpréter que le sens de la règle réside, comme pour une implication logique, autant dans la forme directe que dans la forme contraposée. Dans ce cas, il est légitime de ne pas évaluer la liaison découverte dans les données en tant que simple règle. Plus précisément, derrière la notation $a \rightarrow b$, Kodratoff fait la distinction entre deux types de liaisons de sémantique implicative pouvant être découvertes dans des données [Kod00] :

– Certaines liaisons notées $a \rightarrow b$ expriment pour l'utilisateur une description, comme "les corbeaux sont noirs" (*corbeau* $\rightarrow$ *noir*). Elles doivent être infirmées chaque fois que ($a = 1$ et $b = 0$) est observé, et confirmées chaque fois que ($a = 1$ et $b = 1$) est observé. Comme le suggère le paradoxe de Hempel[9], les cas ($a = 0$ et $b = 0$) qui confirment la contraposition ne sont pas pris en compte.
– Certaines liaisons notées $a \rightarrow b$ expriment pour l'utilisateur une causalité, comme "inhaler de l'amiante provoque le cancer du poumon" (*inhaler_amiante* $\rightarrow$ *cancer_poumon*). Elles doivent être infirmées chaque fois que ($a = 1$ et $b = 0$) est observé, et confirmées chaque fois que ($a = 1$ et $b = 1$) ou ($a = 0$ et $b = 0$) est observé.

Les liaisons qui expriment des descriptions correspondent précisément à la définition des règles donnée dans ce chapitre (tendance de b à être vrai quand a est vrai). En revanche, une liaison qui exprime une causalité n'est pas une règle au sens strict : en considérant les cas ($a = 0$ et $b = 0$) comme des exemples, elle traduit à la fois la règle $a \rightarrow b$ et la règle contraposée $\overline{b} \rightarrow \overline{a}$. Les liaisons

[8] Par exemple avec $(n_{ab}, n_a, n_b, n) = (750; 800; 900; 1000)$, on a $confiance(a \rightarrow b) = 93$ % et $confiance(\overline{b} \rightarrow \overline{a}) = 50$ %.

[9] Le paradoxe de Hempel réside dans le fait qu'un énoncé comme "Tous les corbeaux sont noirs" (logiquement équivalent à "Tout ce qui n'est pas noir n'est pas corbeau") est confirmé par l'observation d'une chaise blanche, d'une chaussure marron, d'un courant d'air...

a \ b	**1**	**0**	
1	exemples n_{ab}	contre-exemples $n_{a\overline{b}}$	n_a
0	$n_{\overline{a}b}$	exemples $n_{\overline{a}\overline{b}}$	$n_{\overline{a}}$
	n_b	$n_{\overline{b}}$	n

TAB. 1.12 – Table de contingence de la quasi-implication $a \Rightarrow b$

exprimant des causalités approximent donc mieux que les règles l'implication logique. Pour cette raison, nous les appellons des *quasi-implications*, notées $a \Rightarrow b$.

Définition 1.6 (liaison exprimant une causalité dans [Kod00]) Une **quasi-implication** est un couple de variables (a, b) noté $a \Rightarrow b$. Les exemples d'une quasi-implication sont les individus de $A \cap B$ et de $\overline{A} \cap \overline{B}$, tandis que les contre-exemples sont les individus de $A \cap \overline{B}$ (voir tableau 1.12). $a \Rightarrow b$ est donc équivalente à sa contraposée $\overline{b} \Rightarrow \overline{a}$. Une quasi-implication est d'autant meilleure qu'elle admet beaucoup d'exemples et peu de contre-exemples.

Nous appelons indice de quasi-implication une mesure de qualité de quasi-implication. Du fait que les quasi-implications expriment pour l'utilisateur une causalité, elles doivent pouvoir être utilisées pour faire des "quasi-déductions" autant dans le sens direct que par contraposition. Les valeurs $I(a \Rightarrow b)$ prises par un indice de quasi-implication doivent donc rendre compte de la qualité des deux règles $a \rightarrow b$ et $\overline{b} \rightarrow \overline{a}$ à la fois : $I(a \Rightarrow b) = I(a \rightarrow b) = I(\overline{b} \rightarrow \overline{a})$.

Définition 1.7 Un **indice de quasi-implication** est un indice de règle I qui vérifie $I(a \rightarrow b) = I(\overline{b} \rightarrow \overline{a})$, c'est-à-dire :

$$I(n_{ab},\ n_a,\ n_b,\ n) = I(n - n_a - n_b + n_{ab},\ n - n_b,\ n - n_a,\ n)$$

C'est sur cette idée de faire valoir la contraposée que Gras [GKCG01] a construit un indice de quasi-implication qui combine explicitement les écarts à l'équilibre de la règle et de la contraposée (par exemple avec la moyenne ou le minimum). Cet indice, appelé indice d'inclusion, est aussi utilisé dans un autre indice de quasi-implication : l'intensité d'implication entropique (voir chapitre 2).

Idéalement, si l'utilisateur est intéressé par des liaisons exprimant des descriptions, alors il est préférable d'employer des indices de règle (au sens strict) et d'éviter les indices de quasi-implication, dont les valeurs sont "parasitées" par la qualité de la contraposée. Au contraire, si l'utilisateur est intéressé par des liaisons exprimant des causalités, alors il vaut mieux employer des indices de quasi-implication et éviter les indices de règle, qui peuvent attribuer à une règle et sa contraposée des valeurs différentes voire même contradictoires. Il est cependant peu probable dans la pratique que l'utilisateur ne soit intéressé dans les données que par une seule de ces deux catégories de règles. A priori,

a \ b	**1**	**0**	
1	exemples n_{ab}	contre-exemples $n_{a\overline{b}}$	n_a
0	contre-exemples $n_{\overline{a}b}$	$n_{\overline{a}\overline{b}}$	$n_{\overline{a}}$
	n_b	$n_{\overline{b}}$	n

(A) Table de contingence de la quasi-conjonction $a \leftrightarrow b$

a \ b	**1**	**0**
1	1	0
0	0	0

(B) Table de vérité de la conjonction logique $a \wedge b$

TAB. 1.13 – Comparaison d'une quasi-conjonction avec la conjonction logique

ce n'est que quand l'utilisateur a lu et interprété une règle qu'il peut juger si sa contraposée est pour lui pertinente ou pas.

Quasi-conjonction

Par analogie avec la quasi-implication qui approxime l'implication logique (qui est une disjonction), nous définissons une nouvelle liaison entre variables qui approxime la conjonction logique : la *quasi-conjonction*.

Définition 1.8 Une **quasi-conjonction** est un couple de variables (a, b) noté $a \leftrightarrow b$. Les exemples d'une quasi-conjonction sont les individus de $A \cap B$, tandis que les contre-exemples sont les individus de $A \cap \overline{B}$ et $\overline{A} \cap B$. $a \leftrightarrow b$ est donc équivalente à sa réciproque $b \leftrightarrow a$. Une quasi-conjonction est d'autant meilleure qu'elle admet beaucoup d'exemples et peu de contre-exemples.

La table de contingence de la quasi-conjonction $a \leftrightarrow b$ et la table de vérité de la conjonction logique $a \wedge b$ sont données dans les tableaux 1.13.(A) et (B). Comme pour la quasi-implication et son homologue logique, les trois quarts des cas sont traités de la même façon :

- les cas ($a = 1$ et $b = 1$) vérifient la quasi-conjonction et la conjonction,
- les cas ($a = 1$ et $b = 0$) et ($a = 0$ et $b = 1$) contredisent la quasi-conjonction et la conjonction,
- seuls les cas ($a = 0$ et $b = 0$) sont considérés différemment : ils contredisent la conjonction mais n'ont pas de rôle défini pour la quasi-conjonction.

Une quasi-conjonction $a \leftrightarrow b$ traduit autant la règle $a \rightarrow b$ que la règle réciproque $b \rightarrow a$. Par analogie avec les indices de quasi-implication, nous proposons d'utiliser comme indices de quasi-conjonction les indices de règle qui évaluent à la fois $a \rightarrow b$ et $b \rightarrow a$.

Définition 1.9 Un **indice de quasi-conjonction** est un indice de règle I qui vérifie $I(a \rightarrow b) = I(b \rightarrow a)$, c'est-à-dire $I(n_{ab},\ n_a,\ n_b,\ n) = I(n_{ab},\ n_b,\ n_a,\ n)$.

a \ b	**1**	**0**	
1	exemples n_{ab}	contre-exemples $n_{a\overline{b}}$	n_a
0	contre-exemples $n_{\overline{a}b}$	exemples $n_{\overline{a}\overline{b}}$	$n_{\overline{a}}$
	n_b	$n_{\overline{b}}$	n

(A) Table de contingence de la quasi-équivalence $a \Leftrightarrow b$

a \ b	**1**	**0**
1	1	0
0	0	1

(B) Table de vérité de l'équivalence logique $a \equiv b$

TAB. 1.14 – Comparaison d'une quasi-équivalence avec l'équivalence logique

Un indice de quasi-conjonction quantifie la concomitance des variables. En particulier, un indice de similarité est un indice de quasi-conjonction, puisqu'un indice de similarité est, d'une part, un indice de règle (voir démonstration 1.1 page 17) et, d'autre part, symétrique par permutation des variables a et b. En revanche, un indice de quasi-conjonction, même positif (comme le lift par exemple), n'est pas un indice de similarité selon la définition qu'en donne Lerman [Ler81].

Si l'utilisateur donne intuitivement du sens à la règle $b \rightarrow a$ quand il lit une règle $a \rightarrow b$ de bonne qualité, il est préférable d'employer des indices de quasi-conjonction plutôt que des indices de règle au sens strict, qui peuvent attribuer à une règle et sa réciproque des valeurs différentes voire même contradictoires. En revanche, si l'utilisateur fait la différence entre une règle et sa réciproque, alors il vaut mieux éviter les indices de quasi-conjonction.

Quasi-équivalence

Définition 1.10 Une **quasi-équivalence** est un couple de variables (a, b) noté $a \Leftrightarrow b$. Les exemples d'une quasi-équivalence sont les individus de $A \cap B$ et $\overline{A} \cap \overline{B}$, tandis que les contre-exemples sont les individus de $A \cap \overline{B}$ et $\overline{A} \cap B$. $a \Leftrightarrow b$ est donc équivalente à sa contraposée $\overline{b} \Leftrightarrow \overline{a}$ et à sa réciproque $b \Leftrightarrow a$. Une quasi-équivalence est d'autant meilleure qu'elle admet beaucoup d'exemples et peu de contre-exemples.

La table de contingence de la quasi-équivalence $a \Leftrightarrow b$ et la table de vérité de l'équivalence logique $a \equiv b$ sont données dans les tableaux 1.14.(A) et (B) (voir aussi [ZZ96]). L'analogie est forte puisque tous les cas sont traités similairement :

- les cas ($a = 1$ et $b = 1$) et ($a = 0$ et $b = 0$) vérifient la quasi-équivalence et l'équivalence,
- les cas ($a = 1$ et $b = 0$) et ($a = 0$ et $b = 1$) contredisent la quasi-équivalence et l'équivalence.

Une quasi-équivalence repose sur les quatre règles $a \rightarrow b$, $b \rightarrow a$, $\overline{b} \rightarrow \overline{a}$, et $\overline{a} \rightarrow \overline{b}$. Nous proposons d'utiliser comme indices de quasi-équivalence les indices

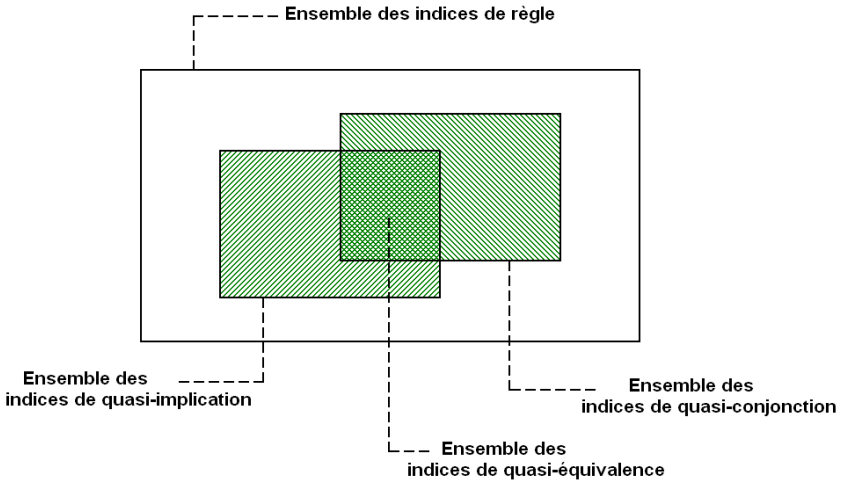


FIG. 1.6 – Les indices de règle selon leur portée

de règle qui évaluent ces quatre règles à la fois.

Définition 1.11 Un **indice de quasi-équivalence** est un indice de règle I qui vérifie $I(a \rightarrow b) = I(b \rightarrow a) = I(\overline{b} \rightarrow \overline{a}) = I(\overline{a} \rightarrow \overline{b})$, c'est-à-dire :

$$\begin{aligned} I(n_{ab},\ n_a,\ n_b,\ n) &= I(n_{ab},\ n_b,\ n_a,\ n) \\ &= I(n - n_a - n_b + n_{ab},\ n - n_b,\ n - n_a,\ n) \\ &= I(n - n_a - n_b + n_{ab},\ n - n_a,\ n - n_b,\ n) \end{aligned}$$

Les indices de quasi-équivalence sont à la fois des indices de quasi-implication et des indices de quasi-conjonction (voir figure 1.6). Ils en possèdent donc les avantages et les inconvénients. Un indice de similarité n'est pas systématiquement un indice de quasi-équivalence. Parmi ceux étudiés dans ce chapitre, seuls les indices de Sokal et Michener et de Rogers et Tanimoto en font partie.

Dans la littérature sur les règles, le terme "équivalence" est souvent associé non pas aux indices de quasi-équivalence mais aux indices de quasi-conjonction. Par exemple, Lerman et Azé [LA04] considèrent que les indices de quasi-conjonction décrits ici évaluent une "similarité d'équivalence $a \Leftrightarrow b$", et Gras *et al.* [GGGP02] définissent une notion de "quasi-équivalence entre variables" qu'ils évaluent avec un "indice entropique d'équivalence" qui est en fait un indice de quasi-conjonction tel que décrit précédemment.

Classification des indices selon la portée

Dans notre classification, la portée d'un indice est l'entité concernée par le résultat de la mesure. Il peut s'agir d'une règle, d'une quasi-implication, d'une quasi-conjonction, ou d'une quasi-équivalence. La classification des indices selon leur portée est donnée dans le tableau 1.15.

Indices de règle	**Indices de quasi-implication**	**Indices de quasi-conjonction**	**Indices de quasi-équivalence**
confiance	indice d'inclusion	lift ou intérêt	coefficient de corrélation
indice de Sebag et Schoenauer	indice de Loevinger	contribution orientée au χ^2	rule-interest
taux exemples et contre-exemples	conviction	indice de vraisemblance du lien	nouveauté
estimateur laplacien	opposé de l'indice d'implication	support ou indice de Russel et Rao	support causal ou indice de Sokal et Michener
indice de Ganascia	intensité d'implication	indice de Jaccard	collective strength
moindre-contradiction		indice de Dice	indice de Yule
multiplicateur de cotes		indice d'Ochiai	rapport de cotes
		indice de Kulczynski	κ
			indice de Rogers et Tanimoto

TAB. 1.15 – Classification des indices de règle selon leur portée

1.4.3 Nature d'un indice de règle

Notre troisième et dernier critère de classification est la nature descriptive ou statistique des indices de règle. Il est aussi recensé dans [LT04] et dans [GCB+04].

Indices descriptifs

Les **indices descriptifs** (ou fréquentiels) ne varient pas avec la dilatation des effectifs (quand tous les effectifs des données sont augmentés ou diminués selon la même proportion). Ils vérifient $I(n_{ab},\ n_a,\ n_b,\ n) = I(\alpha.n_{ab},\ \alpha.n_a,\ \alpha.n_b,\ \alpha.n)$ pour toute constante α strictement positive, et ne permettent donc pas de choisir entre les deux règles $(n_{ab},\ n_a,\ n_b,\ n)$ et $(\alpha.n_{ab},\ \alpha.n_a,\ \alpha.n_b,\ \alpha.n)$[10]. Ces indices prennent en compte les tailles des ensembles d'individus A, B, et $A \cap B$ uniquement de manière relative (par les probabilités $\mathrm{P}(a)$, $\mathrm{P}(b)$, $\mathrm{P}(a \cap b)$) et non de manière absolue (par les effectifs n_a, n_b, n_{ab}). Ce caractère purement descriptif peut sembler pénalisant au premier abord. Toutefois, il est légitime qu'il existe un point de vue selon lequel deux règles $(n_{ab},\ n_a,\ n_b,\ n)$ et $(\alpha.n_{ab},\ \alpha.n_a,\ \alpha.n_b,\ \alpha.n)$ soient considérées de même qualité.

Les indices de règle issus de la théorie de l'information (indices entropiques)

[10] Dans le cadre de l'utilisation d'un indice descriptif avec un seuil de filtrage σ, une façon de neutraliser les effets de la cardinalité des données est de choisir pour σ une fonction $\sigma(n)$.

sont tous des indices descriptifs. Ils sont étudiés au chapitre 2.

Indices statistiques

Les **indices statistiques** sont ceux qui varient avec la dilatation des effectifs. Ils tiennent compte de la taille des phénomènes étudiés, c'est-à-dire que les cardinaux des données sont considérés de manière absolue. Statistiquement, une règle est en effet d'autant plus fiable qu'elle est évaluée sur un grand volume de données. Parmi les indices de règle de nature statistique, il existe une catégorie spécifique : les mesures qui évaluent la significativité statistique des règles. Plus rigoureusement, c'est la significativité statistique de l'écart à l'indépendance ou de l'écart à l'équilibre qui est mesurée.

◇ *Mesures de significativité statistique*

Ces mesures fondées sur un modèle probabiliste comparent la distribution observée des données à une distribution théorique. Dans [Ler81], Lerman développe une méthode de classification hiérarchique nommée AVL (analyse de la vraisemblance des liens). Deux mesures de significativité statistique sont issues de cette méthode :

- l'indice de vraisemblance du lien de Lerman $\mathrm{P}(\mathcal{N}_{ab} < n_{ab})$ [Ler81],
- l'intensité d'implication de Gras $\mathrm{P}(\mathcal{N}_{a\bar{b}} > n_{a\bar{b}})$ [Gra96],

où $\mathcal{N}_{ab}$ et $\mathcal{N}_{a\bar{b}}$ sont les variables aléatoires du nombre d'exemples et du nombre de contre-exemples sous l'hypothèse H_0 d'indépendance entre a et b. Ces indices quantifient respectivement l'invraisemblance de la grandeur du nombre d'exemple n_{ab} et l'invraisemblance de la petitesse du nombre de contre-exemples $n_{a\bar{b}}$, eu égard à l'hypothèse H_0. Bien que chacun de ces indices puisse être interprété comme le complément à 1 de la probabilité critique (*p-value*) d'un test d'hypothèse (unilatéral respectivement à droite et à gauche), il ne s'agit pas ici de tester l'hypothèse d'indépendance H_0 mais bien d'utiliser cette hypothèse comme référence pour évaluer et ordonner les liaisons entre variables.

Lerman [Ler81] propose trois modélisations probabilistes fondées sur différentes hypothèses de tirage pour exprimer H_0. Selon celle qui est retenue, les variables aléatoires suivent soit une loi hypergéométrique, soit une loi binomiale, soit une loi de Poisson. Pour étudier des règles, la modélisation la moins appropriée est la modélisation hypergéométrique. Par ses symétries, elle fait de l'indice de vraisemblance du lien et de l'intensité d'implication un même indice de quasi-équivalence, alors que comme le soulignent Zighed et Rakotomalala [ZR00], "l'étude de la rareté des contre-exemples n'est pas simplement le dual de l'étude de l'abondance des exemples". Nous avons finalement choisi dans ce chapitre (voir tableau 1.3 page 16) de retenir pour l'indice de vraisemblance du lien la modélisation poissonienne, qui est la plus asymétrique [Ler81]. La modélisation poissonienne a aussi l'avantage d'être adaptée à l'étude de variables dont les réalisations sont rares, ce qui convient bien à l'ECD où les bases de données considérées peuvent atteindre de très grandes tailles. Pour l'intensité d'implication, la modélisation communément utilisée est aussi celle de Poisson, qui maximise la dissymétrie par permutation des variables a et b. L'indice de vraisemblance du lien IVL et l'intensité d'implication II s'écrivent alors res-

pectivement :

$$IVL(a \rightarrow b) = \mathrm{P}(Poisson(\frac{n_a n_b}{n}) < n_{ab})$$

$$II(a \rightarrow b) = \mathrm{P}(Poisson(\frac{n_a n_{\overline{b}}}{n}) > n_{a\overline{b}})$$

Avec la modélisation poissonnienne, l'indice de vraisemblance du lien évalue la quasi-conjonction $a \leftrightarrow b$, et l'intensité d'implication évalue la quasi-implication $a \Rightarrow b$. Les deux indices mesurent bien sûr un écart à l'indépendance. Etant donné que ce sont des probabilités, ils ont l'avantage de faire référence à une échelle de valeurs intelligible (échelle de probabilités), ce qui n'est pas le cas de beaucoup d'indices de règle. De tels indices facilitent également la fixation d'un seuil pour le filtrage des règles, puisque si σ est le seuil utilisé pour retenir toutes les règles vérifiant $IVL(a \rightarrow b) > \sigma$ ou $II(a \rightarrow b) > \sigma$, alors $\alpha = 1 - \sigma$ a le sens du risque d'erreur de première espèce d'un test d'hypothèse (on choisit $\alpha \in \{0.1\%, 1\%, 5\%\}$ généralement dans un test).

Par ailleurs, le test le plus communément utilisé pour éprouver l'hypothèse d'indépendance entre deux variables qualitatives a et b est le test d'indépendance du *chi-deux*. La probabilité critique de ce test s'écrit $\mathrm{P}(\chi^2_p > \chi^2_0)$, où la statistique χ^2_p suit la loi du *chi-deux* à p degrés de liberté ($p = 1$ dans le cas de variables binaires), et l'indice de liaison χ^2_0 est calculé sur toute la table de contingence croisant a et b :

$$\chi^2_0 = \frac{(n_{ab} - \frac{n_a n_b}{n})^2}{\frac{n_a n_b}{n}} + \frac{(n_{a\overline{b}} - \frac{n_a n_{\overline{b}}}{n})^2}{\frac{n_a n_{\overline{b}}}{n}} + \frac{(n_{\overline{a}b} - \frac{n_{\overline{a}} n_b}{n})^2}{\frac{n_{\overline{a}} n_b}{n}} + \frac{(n_{\overline{a}\overline{b}} - \frac{n_{\overline{a}} n_{\overline{b}}}{n})^2}{\frac{n_{\overline{a}} n_{\overline{b}}}{n}}$$

Le complément à 1 de cette probabilité critique peut être utilisé comme mesure de la liaison entre a et b [Sap90]. L'avantage de cette mesure est qu'elle ne dépend pas des degrés de liberté (contrairement à l'indice empirique χ^2_0). Cependant, c'est une mesure de liaison entre variables multimodales, puisque χ^2_0 en est elle-même une (voir tableau 1.2 page 12). La probabilité critique du test du *chi-deux* ne différencie pas les exemples des contre-exemples, et il ne peut donc s'agir d'un indice de règle au sens de la définition 1.2.

Si la prise en compte de la taille des phénomènes étudiés fait la force des mesures de significativité, ceci constitue aussi leur principale limite : elles tendent à être peu discriminantes quand les cardinaux étudiés sont grands (de l'ordre de 10^3) [EP96]. En effet, au regard d'effectifs importants, même des écarts triviaux peuvent s'avérer statistiquement significatifs. Dans le cadre de l'évaluation de règles, les deux principales approches pour palier à ce manque de discrimination sont les suivantes :

- Lerman propose d'adapter l'indice de vraisemblance du lien [Ler91] et l'intensité d'implication [LA04] pour qu'ils tiennent compte du "contexte". Plus précisément, les indices sont normalisés vis-à-vis de l'ensemble des valeurs qu'ils prennent sur les données considérées. Ces indices "dans le contexte" ne sont pas étudiés dans ce chapitre car pour mesurer la qualité d'une règle, ils requièrent la totalité de l'ensemble de règles à évaluer. Ce ne sont donc pas des fonctions de $(n_{ab},\ n_a,\ n_b,\ n)$ uniquement.
- Gras propose de moduler l'intensité d'implication par un indice descriptif entropique au sein d'une nouvelle mesure : l'intensité d'implication entropique [GKCG01] [GKB01] [BKGG04]. Elle est présentée en détails

au chapitre suivant.

◇ *Autres indices de nature statistique*

Les mesures de significativité statistique mises à part, il existe des indices de règle qui sont de nature statistique mais ne reposent pas sur un modèle probabiliste. Ces indices sont au nombre de trois (voir formules tableau 1.3 page 16) :

- l'indice *rule-interest* de Piatetsky-Shapiro [PS91],
- la contribution orientée au χ^2 de Lerman [Ler81], notée $q(a \to b)$,
- l'opposé de l'indice d'implication de Gras [Gra96], noté $-ii(a \to b)$ [11].

La contribution orientée au χ^2 et l'indice d'implication interviennent dans les calculs de l'indice de vraisemblance du lien et de l'intensité d'implication quand on fait le choix d'approximer la loi de Poisson par une loi normale. En effet, ils correspondent respectivement à la valeur n_{ab} centrée et réduite selon $\mathcal{N}_{ab}$ et à la valeur $n_{a\overline{b}}$ centrée et réduite selon $\mathcal{N}_{a\overline{b}}$ dans la modélisation poissonienne. Ils sont liés par la relation suivante : $ii(a \to b) = q(a \to \overline{b})$. Ces indices peuvent être interprétés comme une contribution orientée au χ^2 de la table de contingence croisant les variables a et b : $\chi^2 = q(a \to b)^2 + q(a \to \overline{b})^2 + q(\overline{a} \to b)^2 + q(\overline{a} \to \overline{b})^2$ [Ler81]. $q(a \to b)$ est orienté en faveur des exemples, tandis que $ii(a \to b)$ est orienté en faveur des contre-exemples.

1.4.4 Classification

La classification des indices de règle selon l'objet, la portée, et la nature est donnée dans le tableau 1.16. Certaines cellules du tableau sont vides. Tout d'abord, un indice de similarité étant symétrique par permutation des variables, il ne peut s'agir ni d'un indice de règle au sens strict, ni d'un indice de quasi-implication au sens strict. Ensuite, il n'existe aucun indice de quasi-conjonction ou de quasi-équivalence qui mesure un écart à l'équilibre. De tels indices pourraient être développés, mais ils nécessiteraient d'associer des règles dont les équilibres sont différents. Contrairement à l'indépendance, l'équilibre d'une règle $a \to b$ n'est en effet ni l'équilibre de $b \to a$, ni celui de $\overline{b} \to \overline{a}$, ni celui de $\overline{a} \to \overline{b}$. Le seul indice qui combine des équilibres différents (en l'occurrence ceux d'une règle et de sa contraposée) est l'indice d'inclusion.

Alors que l'on considère généralement que les indices de règle sont très nombreux, pour ne pas dire trop, la classification montre qu'il existe en fait peu d'indices de règle au sens strict. En particulier, le seul indice d'écart à l'indépendance qui porte sur des règles au sens strict est le multiplicateur de cotes [LT04]. Par ailleurs, il n'existe aucun indice statistique qui mesure l'écart à l'équilibre. Nous en proposons un dans le chapitre suivant.

1.5 Conclusion

En définissant les notions de *règle* et d'*indice de règle*, ce chapitre établit un cadre formel pour l'étude des règles. Dans ce cadre, nous avons pu comparer

[11] Il est nécessaire de prendre l'opposé pour que les fortes valeurs soient associées aux règles admettant peu de contre-exemples, et ainsi obtenir un indice de règle.

les règles aux concepts connexes que sont les similarités, les implications, et les équivalences. Nous avons également réalisé une classification inédite des principaux indices de règle de la littérature selon trois critères : l'objet, la portée, et la nature.

- L'objet est la notion qui est mesurée par l'indice. Il peut s'agir d'un écart à l'équilibre, d'un écart à l'indépendance, ou plus anecdotiquement d'une similarité. Ecart à l'équilibre et écart à l'indépendance sont deux aspects différents mais complémentaires de la qualité des règles.
- La portée est l'entité concernée par le résultat de la mesure. Il peut s'agir d'une unique règle, ou bien d'une règle et de sa contraposée (quasi-implication), ou bien d'une règle et de sa réciproque (quasi-conjonction), ou bien d'une règle et de sa contraposée et de sa réciproque (quasi-équivalence).
- La nature est le caractère descriptif ou statistique de l'indice.

Ces trois critères nous paraissent essentiels pour appréhender la signification des indices. Ainsi, la classification permet d'aider l'utilisateur à choisir quels indices appliquer pour valider les règles. Elle amène par exemple à se demander si l'utilisateur s'intéresse uniquement à des règles au sens strict, ou bien si la contraposée et la réciproque peuvent faire sens pour lui. Il est également pertinent de se demander si l'utilisateur désire mesurer des écarts à l'équilibre ou des écarts à l'indépendance, ou bien les deux. En l'absence d'indication de la part de l'utilisateur, il nous paraît judicieux d'employer conjointement un indice descriptif d'écart à l'équilibre, un indice statistique d'écart à l'équilibre, un indice descriptif d'écart à l'indépendance, et un indice statistique d'écart à l'indépendance. Selon nous, un tel quadruplet d'indices permet de mesurer quatre aspects fortement "orthogonaux" de la qualité des règles.

Portée / **Objet**	**Règle**	**Quasi-implication**	**Quasi-conjonction**	**Quasi-équivalence**
Ecart à l'équilibre	– confiance, – indice de Sebag et Schoenauer, – taux des exemples et contre-exemples, – estimateur laplacien de probabilité conditionnelle, – indice de Ganascia, – moindre-contradiction	– indice d'inclusion		
Ecart à l'indépendance	– multiplicateur de cotes	– indice de Loevinger, – conviction – *intensité d'implication*, – *indice d'implication*	– lift ou intérêt – *indice de vraisemblance du lien*, – *contribution orientée au* χ^2	– coefficient de corrélation, – nouveauté, – collective strength, – κ, – indice de Yule, – rapport de cotes – *rule-interest*
Similarité			– support ou indice de Russel et Rao, – indice de Jaccard, – indice de Dice, – indice d'Ochiai, – indice de Kulczynski	– support causal ou indice de Sokal et Michener, – indice de Rogers et Tanimoto

La **nature** des indices est indiquée par le style de la police : les indices en *italique* sont statistiques, les autres sont descriptifs.

Tab. 1.16 – Classification des indices de règle

Trois indices de règle : IPEE, intensité d'implication entropique, taux informationnel

2

Sommaire

Dans ce chapitre, nous présentons trois nouveaux indices de règle : IPEE, l'intensité d'implication entropique, et le taux informationnel. Ils possèdent tous les trois des caractéristiques originales, c'est pourquoi nous leur consacrons un chapitre à part entière. Le premier indice, nommé IPEE, est le seul indice d'écart à l'équilibre qui soit de nature statistique. Il est fondé sur un modèle probabiliste et évalue la significativité de l'écart à l'équilibre. L'intensité d'implication entropique, quant à elle, est une extension de l'intensité d'implication mieux adaptée aux grands jeux de données. Elle prend en compte à la fois l'écart à l'équilibre et l'écart à l'indépendance. Enfin, le taux informationnel est une mesure fondée sur la théorie de l'information. Elle possède la particularité unique de rejeter simultanément les mauvais écarts à l'équilibre et les mauvais écarts à l'indépen-

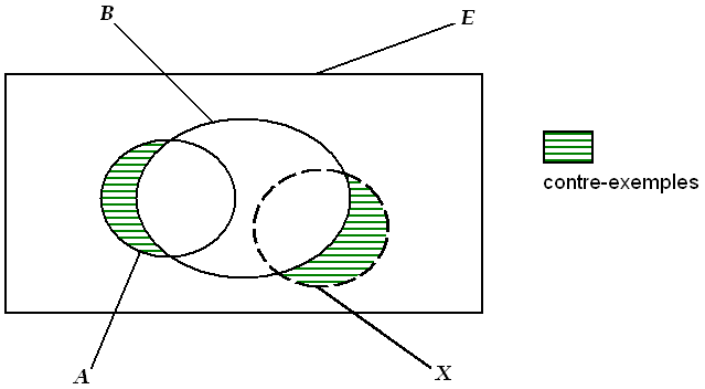


FIG. 2.1 – Tirage aléatoire d'un ensemble X sous hypothèse d'équiprobabilité entre les exemples et les contre-exemples

dance. Pour chacun des trois indices de règle, nous décrivons sa construction et étudions ses propriétés.

2.1 IPEE, un indice probabiliste d'écart à l'équilibre

Nous avons vu au chapitre précédent qu'il n'existe aucun indice statistique qui mesure l'écart à l'équilibre. Pourtant, les indices statistiques ont l'avantage de prendre en compte les cardinaux des données de manière absolue. Statistiquement, une règle est en effet d'autant plus fiable qu'elle est évaluée sur un grand volume de données. De plus, un indice statistique, lorsqu'il est fondé sur un modèle probabiliste, fait référence à une échelle de valeurs intelligible (échelle de probabilités), ce qui n'est pas le cas de beaucoup d'indices de règle. Un tel indice facilite également la fixation d'un seuil pour le filtrage des règles, puisque le complément à 1 du seuil a le sens du risque d'erreur de première espèce d'un test d'hypothèse (on choisit $\alpha \in \{0.1\%, 1\%, 5\%\}$ généralement dans un test). Pour ces différentes raisons, nous proposons un nouvel indice de règle qui mesure l'écart à l'équilibre tout en étant de nature statistique. Plus précisément, cet indice évalue la significativité de l'écart à l'équilibre, là où l'intensité d'implication ou l'indice de vraisemblance du lien évaluent la significativité de l'écart à l'indépendance.

2.1.1 Modèle aléatoire

Etant donnée une règle $a \rightarrow b$, nous cherchons à mesurer la significativité statistique de l'écart à l'équilibre de la règle. La configuration d'équilibre étant définie par l'équirépartition dans A des exemples $A \cap B$ et des contre-exemples $A \cap \overline{B}$, l'hypothèse de référence est l'hypothèse H_0 d'équiprobabilité entre les exemples et les contre-exemples. Associons donc à l'ensemble A un ensemble aléatoire X de cardinal n_a tiré dans E sous cette hypothèse : $\mathrm{P}(X \cap B) =$

$\mathrm{P}(X \cap \overline{B})$ (voir figure 2.1). Le nombre de contre-exemples attendu sous H_0 est le cardinal de $X \cap \overline{B}$. Il s'agit d'une variable aléatoire, notée $\mathcal{N}_{a\bar{b}}$, dont $n_{a\bar{b}}$ est une valeur observée. La règle $a \rightarrow b$ est d'autant meilleure que la probabilité que le hasard produise plus de contre-exemples que les données est grande.

Définition 2.1 L'**indice probabiliste d'écart à l'équilibre (IPEE)** d'une règle $a \rightarrow b$ est défini par :

$$IPEE(a \rightarrow b) = \mathrm{P}(\mathcal{N}_{a\bar{b}} > n_{a\bar{b}} \mid \mathrm{H}_0)$$

Une règle $a \rightarrow b$ est dite admissible au seuil de confiance $1 - \alpha$ si $IPEE(a \rightarrow b) \geq 1 - \alpha$.

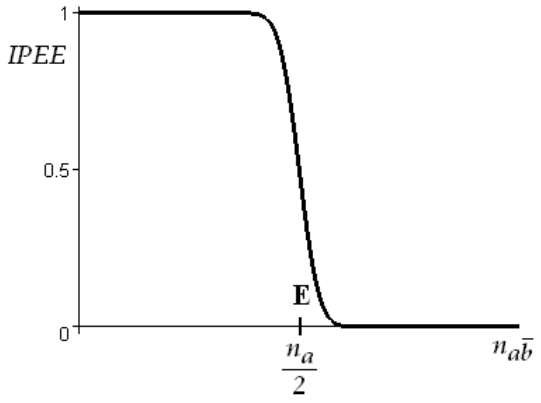


FIG. 2.2 – Représentation de IPEE en fonction de $n_{a\bar{b}}$
(E : équilibre)

IPEE quantifie donc l'invraisemblance de la petitesse du nombre de contre-exemples $n_{a\bar{b}}$ eu égard à l'hypothèse H_0. En particulier :

- si $IPEE(a \rightarrow b)$ vaut 0, alors il est invraisemblable que les caractères (a *et* b) et (a *et* $\bar{b}$) soient équiprobables (l'écart à l'équilibre de la règle est significatif mais orienté en faveur des contre-exemples) ;
- si $IPEE(a \rightarrow b)$ vaut 0.5, alors il est vraisemblable que les caractères (a *et* b) et (a *et* $\bar{b}$) soient équiprobables (l'écart à l'équilibre de la règle n'est pas significatif) ;
- si $IPEE(a \rightarrow b)$ vaut 1, alors il est invraisemblable que les caractères (a *et* b) et (a *et* $\bar{b}$) soient équiprobables (l'écart à l'équilibre de la règle est significatif et orienté en faveur des exemples) ;

Ce nouvel indice peut être interprété comme le complément à 1 de la probabilité critique (*p-value*) d'un test d'hypothèse (et α comme le risque de première espèce de ce test). Toutefois, à l'instar de l'intensité d'implication et de l'indice de vraisemblance du lien (où H_0 est l'hypothèse d'indépendance entre a et b), il ne s'agit pas ici de tester une hypothèse mais bien de l'utiliser comme référence pour évaluer et ordonner les règles.

2.1.2 Expression analytique

Dans le cadre d'un tirage avec remise, $\mathcal{N}_{a\bar{b}}$ suit une loi binomiale de paramètres $\frac{1}{2}$ (autant de chances de tirer un exemple que de tirer un contre-exemple) et n_a. IPEE s'écrit donc :

$$IPEE(a \to b) = 1 - \frac{1}{2^{n_a}} \sum_{k=0}^{n_{a\bar{b}}} \mathrm{C}_{n_a}^{k}$$

IPEE ne dépend ni de n_b, ni de n puisque l'hypothèse d'équilibre H_0 ne se définit pas à l'aide de n_b et de n (contrairement à l'hypothèse d'indépendance). Il est à noter que la significativité statistique de l'écart à l'équilibre pourrait aussi être mesurée en comparant non pas les contre-exemples mais les exemples : $\widehat{IPEE}(a \to b) = \mathrm{P}(\mathcal{N}_{ab} < n_{ab} \mid \mathrm{H}_0)$. Cependant, les distributions binomiales de paramètre $\frac{1}{2}$ étant symétriques, les deux indices sont identiques :

$$IPEE(a \to b) = 1 - \frac{1}{2^{n_a}} \sum_{K=n_{ab}}^{n_a} \mathrm{C}_{n_a}^{n_a-K} = 1 - \frac{1}{2^{n_a}} \sum_{K=n_{ab}}^{n_a} \mathrm{C}_{n_a}^{K} = \widehat{IPEE}(a \to b)$$

$$\text{où } K = n_a - k$$

Quand $n_a > 15$, la loi binomiale peut être approximée par une loi normale, et IPEE s'écrit alors :

$$IPEE(a \to b) = \mathrm{P}(\mathcal{N}(0,1) > \widetilde{n}_{a\bar{b}}) = \frac{1}{\sqrt{2\pi}} \int_{\widetilde{n}_{a\bar{b}}}^{\infty} e^{-\frac{t^2}{2}} \mathrm{dt}$$

où la variable aléatoire $\mathcal{N}(0,1)$ suit la loi normale centrée réduite,

$$\text{et } \widetilde{n}_{a\bar{b}} = \frac{n_{a\bar{b}} - \frac{n_a}{2}}{\sqrt{\frac{n_a}{4}}} \text{ est la valeur observée.}$$

L'effectif centré réduit $\widetilde{n}_{a\bar{b}}$ peut être être interprété comme la contribution orientée au χ^2 d'adéquation entre la distribution observée exemples/contre-exemples et la distribution uniforme : $\chi^2 = \widetilde{n}_{a\bar{b}}{}^2$. Ceci constitue une analogie forte avec l'intensité d'implication et l'indice de vraisemblance du lien, puisque dans la modélisation poissonnienne associée à ces indices, les valeurs centrées réduites de $n_{a\bar{b}}$ et n_{ab} peuvent être interprétées comme des contributions orientées au χ^2 d'indépendance entre a et de b [Ler81].

2.1.3 Propriétés

IPEE est un indice de règle au sens de la définition 1.2. Ses principales propriétés sont décrites dans le tableau 2.1. L'indice est représenté en fonction du nombre de contre-exemples dans la figure 2.2, et comparé dans la figure 2.3 aux principaux indices d'écarts à l'équilibre : la confiance, la moindre-contradiction, et l'indice d'inclusion. Nous pouvons voir que :

- IPEE réagit faiblement aux premiers contre-exemples (décroissance lente). Ce comportement est intuitivement satisfaisant pour un indice statistique puisqu'un faible nombre de contre-exemples ne saurait remettre en cause la règle [GCB+04].

Objet	écart à l'équilibre
Portée	règle au sens strict
Nature	statistique
Domaine de variation	$[0\ ;\ 1]$
Valeur pour les règles logiques	$1 - \frac{1}{2^{n_a}}$
Valeur pour les règles à l'équilibre	0.5
Valeur pour les règles à l'indépendance	< 1

TAB. 2.1 – Propriétés de IPEE

– Le rejet des règles s'accélère dans une zone d'incertitude autour de l'équilibre $n_{a\bar{b}} = \frac{n_a}{2}$ (décroissance rapide).

Dans les figures 2.4, les effectifs des données sont multipliés par un coefficient γ à partir d'une configuration initiale. Les indices sont représentés en fonction de γ. Ces figures montrent qu'à proportion exemples/contre-exemples fixée, les indices sont constants sauf IPEE dont les valeurs sont d'autant plus extrêmes (proches de 0 ou 1) que n_a est grand[1]. En effet, de par sa nature statistique, l'indice prend en compte la taille des phénomènes étudiés : plus n_a est grand, plus on peut avoir confiance dans le déséquilibre exemples/contre-exemples observé dans les données, et plus on peut confirmer (figure 2.4.(A)) ou infirmer (figure 2.4.(B)) la bonne qualité de la règle. En particulier, pour IPEE, la qualité d'une règle logique dépend de n_a (voir tableau 2.1). Ainsi IPEE a l'avantage de ne pas attribuer systématiquement la même valeur aux règles logiques. Ceci permet de différencier et hiérarchiser les règles logiques. Parmi les indices d'écart à l'équilibre (voir tableau 1.16 page 39), seuls la moindre-contradiction et IPEE possèdent cette caractéristique : la moindre-contradiction différencie les règles logiques selon n_b (l'indice favorise les conclusions rares), tandis que IPEE différencie les règles logiques selon n_a (l'indice favorise les prémisses fréquentes).

IPEE porte sur des règles au sens strict et ne possède donc aucune symétrie. On a toutefois la relation suivante :

$$IPEE(a \rightarrow \bar{b}) = 1 - IPEE(a \rightarrow b) - \frac{\mathrm{C}_{n_a}^{n_{ab}}}{2^{n_a}}$$

(le dernier terme est négligeable quand n_a est grand)

Comme nous l'avons vu au chapitre 1, les mesures de significativité statistique tendent à être peu discriminantes quand les cardinaux étudiés sont grands (de l'ordre de 10^3), car même des écarts triviaux peuvent s'avérer statistiquement significatifs au regard d'effectifs importants. Comme l'illustre la figure 2.5, IPEE ne déroge pas à la règle : quand n_a est grand, l'indice tend à évaluer que les règles sont soit très bonnes (valeurs proches de 1), soit très mauvaises (valeurs

[1]Quand la modélisation retenue est gaussienne, ce comportement est visible directement sur $\widetilde{n}_{a\bar{b}}$: $\widetilde{n}_{a\bar{b}} = \sqrt{n_a}(1 - 2 \times confiance)$.

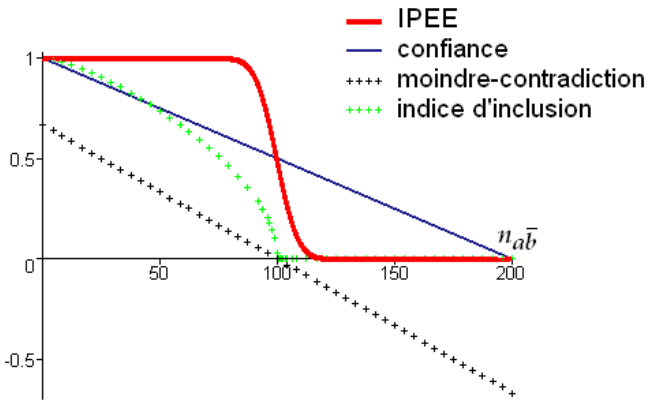
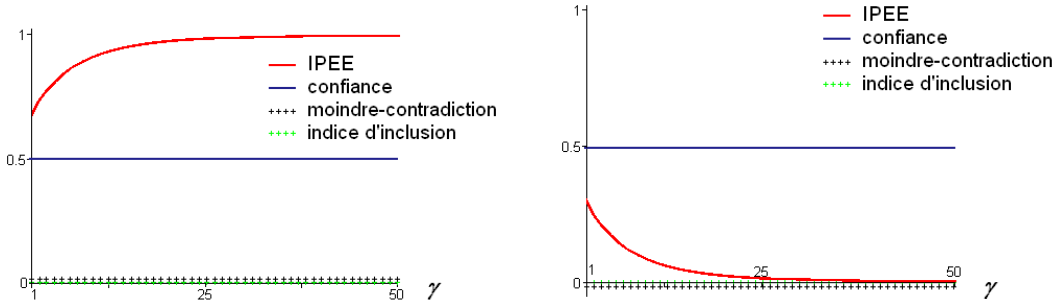


FIG. 2.3 – Représentation des indices d'écart à l'équilibre en fonction de $n_{a\bar{b}}$ ($n_a = 200,\ n_b = 300,\ n = 1000,\ n_{a\bar{b}} \in [0\ ;\ 200]$)



(A) $n_a = 200 \times \gamma,\ n_b = 400 \times \gamma,$ $n_{a\bar{b}} = 97 \times \gamma,\ n = 1000 \times \gamma,$ $\gamma \in [1\ ,\ 50]$

(B) $n_a = 200 \times \gamma,\ n_b = 400 \times \gamma,$ $n_{a\bar{b}} = 103 \times \gamma,\ n = 1000 \times \gamma,$ $\gamma \in [1\ ,\ 50]$

FIG. 2.4 – Représentation des indices d'écarts à l'équilibre en fonction de la dilatation des effectifs

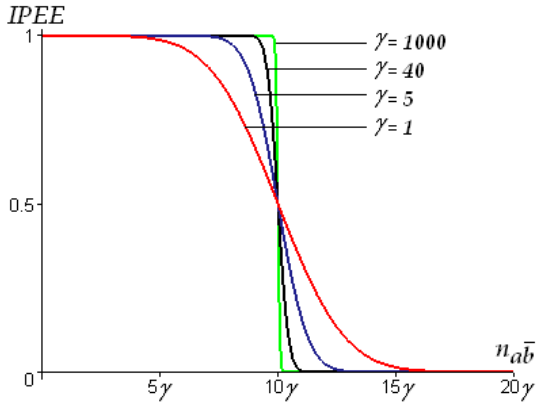


FIG. 2.5 – Représentation de IPEE avec la dilatation des effectifs ($n_a = 20 \times \gamma,\ n_{a\bar{b}} \in [0 \times \gamma\ ;\ 20 \times \gamma],\ \gamma \in \{1;\ 5;\ 40;\ 1000\}$)

proches de 0). Dans ce cas, pour affiner le filtrage des meilleures règles, il faut utiliser en supplément de IPEE une mesure descriptive. En revanche, contrairement à l'intensité d'implication ou à l'indice de vraisemblance du lien, IPEE ne dépend pas de n. L'indice est donc autant sensible aux règles spécifiques ("pépites de connaissances") qu'aux règles générales, et a l'avantage d'être adapté à l'étude des petites bases de données comme des grandes.

2.2 L'intensité d'implication entropique

L'*intensité d'implication* est un indice de quasi-implication développé par Gras [Gra96] et qui est au fondement d'une méthode d'analyse exploratoire de données nommée *analyse statistique implicative* [GKB01]. Cet indice quantifie l'invraisemblance de la petitesse du nombre de contre-exemples $n_{a\bar{b}}$ eu égard à l'hypothèse d'indépendance entre a et b. Comme toutes les mesures de significativité statistique, cet indice est peu discriminant quand les cardinaux étudiés sont grands (voir chapitre 1). Pour résoudre ce problème, Gras *et al.* ont proposé dans [GKCG01] de moduler les valeurs de l'intensité d'implication par un indice de quasi-implication descriptif fondé sur l'entropie de Shannon : l'*indice d'inclusion*. Le nouvel indice ainsi formé s'appelle *intensité d'implication entropique*. Il a la particularité de prendre en compte à la fois l'écart à l'équilibre et l'écart à l'indépendance.

2.2.1 Rappels sur l'intensité d'implication

Modèle aléatoire

Pour une règle $a \rightarrow b$, l'intensité d'implication compare le nombre de contre-exemples $n_{a\bar{b}}$ observé dans les données au nombre de contre-exemples attendu sous l'hypothèse H_0 d'indépendance entre a et b. Associons donc aux ensembles

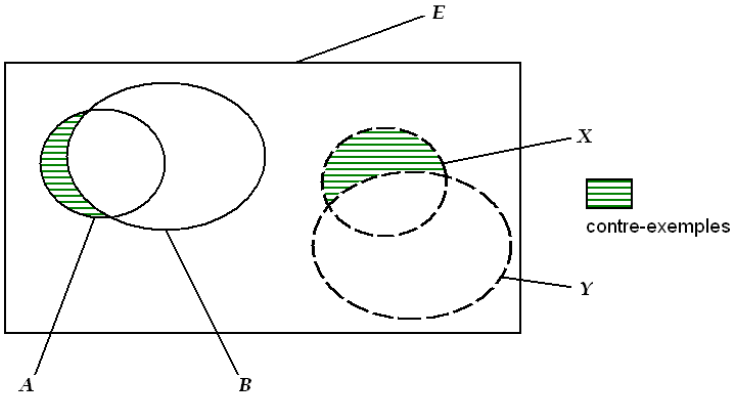


FIG. 2.6 – Tirage aléatoire de deux ensembles indépendants X et Y

A et B deux ensembles indépendants X et Y tirés aléatoirement dans E et de mêmes cardinaux que A et B : $|X| = n_a$ et $|Y| = n_b$ (figure 2.6). Le nombre de contre-exemples attendu sous H_0 est le cardinal de $X \cap \overline{Y}$. Il s'agit d'une variable aléatoire, notée $\mathcal{N}_{a\bar{b}}$, dont $n_{a\bar{b}}$ est une valeur observée. La règle $a \rightarrow b$ est d'autant meilleure que la probabilité que le hasard produise plus de contre-exemples que les données est grande.

Définition 2.2 L'**intensité d'implication** φ d'une règle $a \rightarrow b$ est définie par :

$$\varphi(a \rightarrow b) = \mathrm{P}(\mathcal{N}_{a\bar{b}} > n_{a\bar{b}} \mid \mathrm{H}_0)$$

Une règle $a \rightarrow b$ est dite admissible au seuil de confiance $1-\alpha$ si $\varphi(a \rightarrow b) \geq 1-\alpha$.

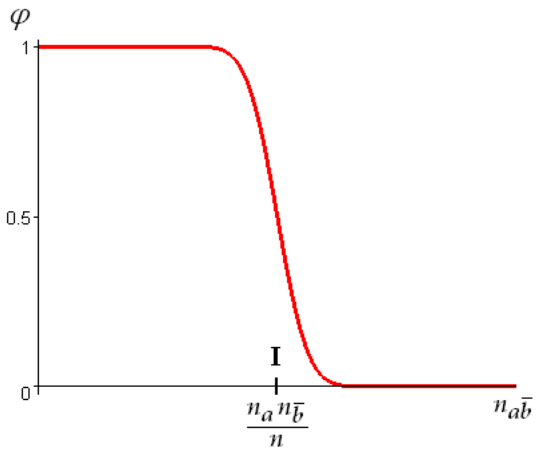


FIG. 2.7 – Représentation de l'intensité d'implication en fonction de $n_{a\bar{b}}$ (I : indépendance)

Expression analytique

Lerman a montré dans [Ler81] qu'en fonction des hypothèses de tirage retenues, la variable aléatoire $\mathcal{N}_{a\bar{b}}$ suit soit la loi hypergéométrique de paramètres $(n, n_a, n_{\bar{b}})$, soit la loi binomiale de paramètres $(n, \frac{n_a n_{\bar{b}}}{n^2})$, soit la loi de Poisson de paramètre $\frac{n_a n_{\bar{b}}}{n}$. Comme indiqué au chapitre 1 page 35, nous choisissons dans cette thèse la distribution de Poisson. En posant $\lambda = \frac{n_a n_{\bar{b}}}{n}$, l'intensité d'implication s'écrit :

$$\varphi(a \rightarrow b) = 1 - \sum_{k=0}^{n_{a\bar{b}}} \frac{\lambda^k}{k!} e^{-\lambda}$$

Quand $\lambda > 15$, la loi de Poisson peut être approximée par une loi normale, et l'intensité d'implication s'écrit alors :

$$\varphi(a \rightarrow b) = \mathrm{P}(\mathcal{N}(0,1) > \widetilde{n}_{a\bar{b}}) = \frac{1}{\sqrt{2\pi}} \int_{\widetilde{n}_{a\bar{b}}}^{\infty} e^{-\frac{t^2}{2}} \mathrm{dt}$$

où la variable aléatoire $\mathcal{N}(0,1)$ suit la loi normale centrée réduite

$$\text{et } \widetilde{n}_{a\bar{b}} = \frac{n_{a\bar{b}} - \lambda}{\sqrt{\lambda}} \text{ est la valeur observée.}$$

$\widetilde{n}_{a\bar{b}}$ est appelé indice d'implication [Gra96]. Son opposé est un indice de règle (voir chapitre 1).

2.2.2 L'indice d'inclusion, un indice descriptif fondé sur l'entropie

Gras a développé l'indice d'inclusion spécialement pour l'associer à l'intensité d'implication. Afin que l'association soit cohérente, il fallait que l'indice d'inclusion évalue des quasi-implications, comme l'intensité d'implication. Pour cela, Gras a fondé l'indice d'inclusion sur deux écarts à l'équilibre :

- l'écart à l'équilibre de la règle à évaluer $a \rightarrow b$, associé au déséquilibre entre les exemples $A \cap B$ et les contre-exemples $A \cap \overline{B}$,
- l'écart à l'équilibre de la règle contraposée $\overline{b} \rightarrow \overline{a}$, associé au déséquilibre entre les exemples $\overline{A} \cap \overline{B}$ et les contre-exemples $A \cap \overline{B}$.

Comme nous l'avons vu dans le chapitre 1, l'équilibre d'une règle n'est pas l'équilibre de sa contraposée (contrairement à l'indépendance qui, elle, est commune aux deux règles).

Une mesure bien connue pour évaluer les déséquilibres de façon non linéaire est l'entropie de Shannon [SW49]. Considérons l'expérience aléatoire qui consiste à vérifier si b est vrai quand a est vrai. L'incertitude moyenne de l'expérience est donnée par l'entropie conditionnelle $H(b \ / \ a = 1)$ de la variable b sachant la réalisation de a [2] :

$$H(b \ / \ a = 1) = -\frac{n_{ab}}{n_a} \log_2 \frac{n_{ab}}{n_a} - \frac{n_{a\bar{b}}}{n_a} \log_2 \frac{n_{a\bar{b}}}{n_a}$$

[2] Les fonctions entropiques associent des variables et des réalisations de variables. Pour plus de clarté, les réalisations d'une variable booléenne a sont notées $a = 1$ et $a = 0$ dans les fonctions entropiques, et non pas a et $\overline{a}$ comme dans les autres notations.

Similairement, l'entropie conditionnelle $H(\overline{a} \ / \ b = 0)$ quantifie l'incertitude moyenne de l'expérience aléatoire qui consiste à vérifier si a est faux quand b est faux :

$$H(\overline{a} \ / \ b = 0) = -\frac{n_{\overline{a}\overline{b}}}{n_{\overline{b}}} \log_2 \frac{n_{\overline{a}\overline{b}}}{n_{\overline{b}}} - \frac{n_{a\overline{b}}}{n_{\overline{b}}} \log_2 \frac{n_{a\overline{b}}}{n_{\overline{b}}}$$

L'indice d'inclusion utilise les entropies conditionnelles $H(b/a = 1)$ et $H(\overline{a}/b = 0)$ pour mesurer les écarts à l'équilibre d'une règle et de sa contraposée. Des règles de bonne qualité du point de vue de l'écart à l'équilibre engendrent des entropies faibles. Pour obtenir une mesure unique, les entropies sont combinées par la moyenne géométrique :

$$\sqrt{(1 - H(b/a = 1))(1 - H(\overline{a}/b = 0))}$$

Les compléments à 1 permettent d'associer les règles de bonne qualité aux valeurs fortes et non aux valeurs faibles. Pour renforcer le contraste entre les petites et les grandes entropies, celles-ci sont élevées à la puissance d'un nombre réel fixé $\omega \geq 1$:

$$\sqrt[2\omega]{(1 - H(b/a = 1)^{\ \omega})(1 - H(\overline{a}/b = 0)^{\ \omega})}$$

Par leur symétrie, les entropies conditionnelles $H(b/a = 1)$ et $H(\overline{a}/b = 0)$ évaluent identiquement un déséquilibre en faveur des exemples et un déséquilibre en faveur des contre-exemples : $H(b/a = 1) = H(\overline{b}/a = 1)$ et $H(\overline{a}/b = 0) = H(a/b = 0)$. Afin d'obtenir un indice de règle selon la définition 1.2, seul le déséquilibre en faveur des exemples doit être retenu. Pour cela, comme dans [GKB01] et [BKGG04], nous considérons qu'une règle est sans intérêt lorsque les contre-exemples sont plus nombreux que les exemples, et annulons les termes $1 - H(b/a = 1)^{\ \omega}$ et $1 - H(\overline{a}/b = 0)^{\ \omega}$ quand $n_{a\overline{b}} \geq \frac{n_a}{2}$ et $n_{a\overline{b}} \geq \frac{n_{\overline{b}}}{2}$ respectivement. Une autre solution est proposée dans [GKCG01] mais elle est moins intuitive et engendre un indice de règle moins filtrant.

Définition 2.3 L'**indice d'inclusion** τ d'une règle $a \rightarrow b$ est défini par :

$$\tau(a \rightarrow b) = \sqrt[2\omega]{I^{\omega}_{b/a=1} I^{\omega}_{\overline{a}/b=0}} \qquad \text{où } \omega \geq 1 \text{ et :}$$

$$I^{\omega}_{b/a=1} = \begin{cases} 1 - \left(-\frac{n_{ab}}{n_a} \log_2 \frac{n_{ab}}{n_a} - \frac{n_{a\overline{b}}}{n_a} \log_2 \frac{n_{a\overline{b}}}{n_a} \right)^{\omega} & \text{si } n_{a\overline{b}} < \frac{n_a}{2} \\ 0 & \text{sinon} \end{cases}$$

$$I^{\omega}_{\overline{a}/b=0} = \begin{cases} 1 - \left(-\frac{n_{\overline{a}\overline{b}}}{n_{\overline{b}}} \log_2 \frac{n_{\overline{a}\overline{b}}}{n_{\overline{b}}} - \frac{n_{a\overline{b}}}{n_{\overline{b}}} \log_2 \frac{n_{a\overline{b}}}{n_{\overline{b}}} \right)^{\omega} & \text{si } n_{a\overline{b}} < \frac{n_{\overline{b}}}{2} \\ 0 & \text{sinon} \end{cases}$$

L'indice d'inclusion présente l'originalité d'être le seul indice d'écart à l'équilibre qui porte sur des quasi-implications (voir classification 1.16 page 39). Il s'annule dès que l'écart à l'équilibre de la règle ou de sa contraposée n'est pas orienté en faveur des exemples, c'est-à-dire lorsque $n_{a\overline{b}} \geq \min(\frac{n_a}{2}, \frac{n_{\overline{b}}}{2})$.

ω est un paramètre de sélectivité de l'indice d'inclusion qui peut être ajusté en fonction des données étudiées : plus ω est faible, plus l'indice d'inclusion décroît rapidement avec les contre-exemples, et plus le filtrage des règles est sévère (figure 2.8). En analyse statistique implicative, c'est généralement $\omega = 2$ qui est retenu. Ce choix engendre un indice d'inclusion qui réagit faiblement aux

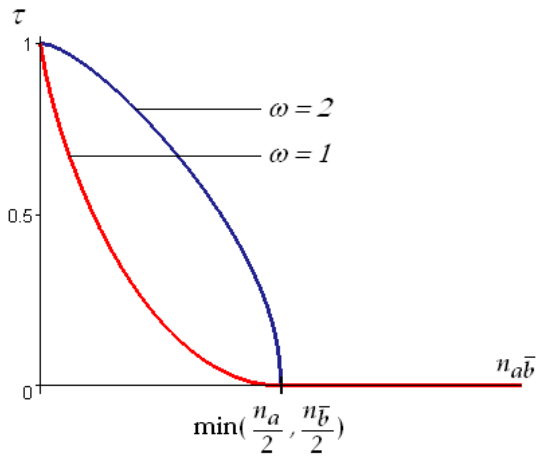


FIG. 2.8 – Représentation de l'indice d'inclusion en fonction de $n_{a\bar{b}}$

premiers contre-exemples, ce qui est pour Gras une proriété fondamentale d'un bon indice de règle [GCB+04]. Comme dans le chapitre 1, nous choisissons dans la suite $\omega = 2$.

2.2.3 Association des deux indices

L'association de l'intensité d'implication et de l'indice d'inclusion crée un indice de quasi-implication, nommé *intensité d'implication entropique*, qui est de nature statistique (grâce à l'intensité d'implication) tout en restant discriminant quand les cardinaux étudiés sont grands (grâce à l'indice d'inclusion). L'association des deux mesures est réalisée par la moyenne géométrique [GKCG01].

Définition 2.4 L'**intensité d'implication entropique** ϕ d'une règle $a \rightarrow b$ est définie par :

$$\phi(a \rightarrow b) = \sqrt{\varphi(a \rightarrow b) \times \tau(a \rightarrow b)}$$

2.2.4 Propriétés

Intensité d'implication

Les principales propriétés de l'intensité d'implication sont résumées dans le tableau 2.2. Dans la figure 2.9, l'indice est représenté en fonction du nombre de contre-exemples et comparé aux principaux indices d'écarts à l'indépendance : l'indice de Loevinger, le lift, et la corrélation. Nous pouvons voir que :

- L'intensité d'implication réagit faiblement aux premiers contre-exemples (décroissance lente). Ce comportement est intuitivement satisfaisant pour un indice statistique puisqu'un faible nombre de contre-exemples ne saurait remettre en cause la règle [GCB+04].

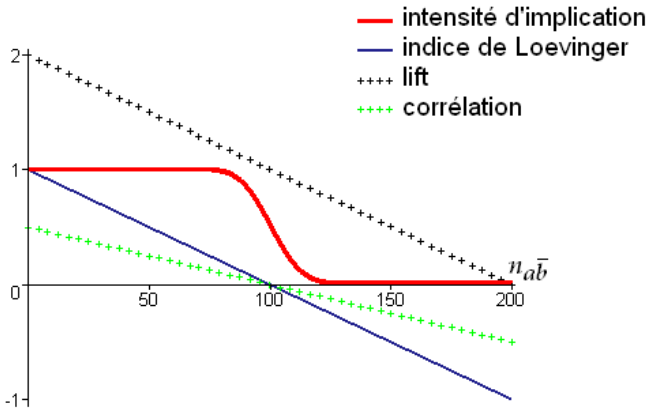
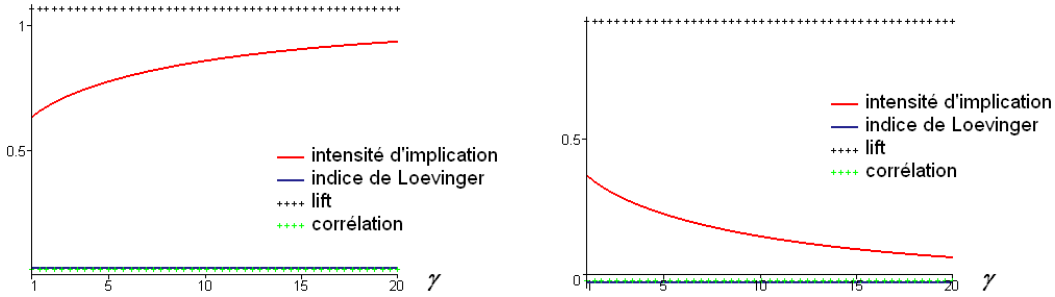


FIG. 2.9 – Représentation des indices d'écart à l'indépendance selon $n_{a\bar{b}}$
($n_a = 200,\ n_b = 500,\ n = 1000,\ n_{a\bar{b}} \in [0\ ;\ 200]$)



(A) $n_a = 200 \times \gamma,\ n_b = 300 \times \gamma,$
$n_{a\bar{b}} = 136 \times \gamma,\ n = 1000 \times \gamma,$
$\gamma \in [1\ ,\ 20]$

(B) $n_a = 200 \times \gamma,\ n_b = 300 \times \gamma,$
$n_{a\bar{b}} = 144 \times \gamma,\ n = 1000 \times \gamma,$
$\gamma \in [1\ ,\ 20]$

FIG. 2.10 – Représentation des indices d'écarts à l'indépendance en fonction de la dilatation des effectifs

Objet	écart à l'indépendance
Portée	quasi-implication
Nature	statistique
Domaine de variation	$[0\ ;\ 1]$
Valeur pour les règles logiques	$1 - e^{-\frac{n_a n_{\bar{b}}}{n}}$
Valeur pour les règles à l'équilibre	< 1
Valeur pour les règles à l'indépendance	0

TAB. 2.2 – Propriétés de l'intensité d'implication

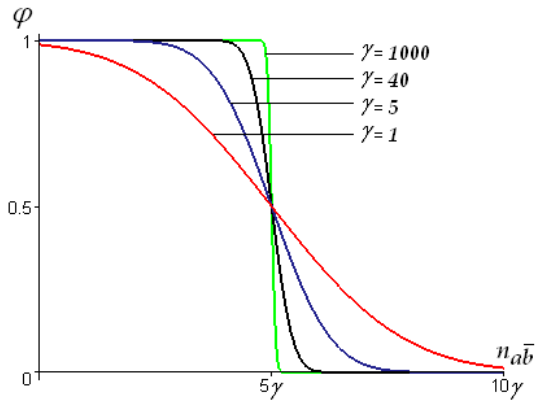


FIG. 2.11 – Représentation de l'intensité d'implication avec la dilatation des effectifs

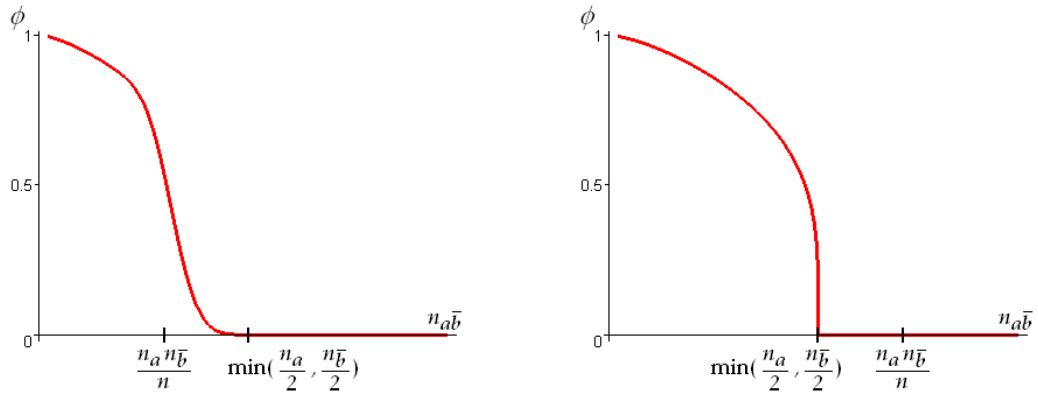
$$(n_a = 20 \times \gamma,\ n_b = 75 \times \gamma,\ n = 100 \times \gamma,$$
$$n_{a\bar{b}} \in [0 \times \gamma\ ;\ 10 \times \gamma],\ \gamma \in \{1;\ 5;\ 40;\ 1000\})$$

Objet	écart à l'équilibre
Portée	quasi-implication
Nature	statistique
Domaine de variation	$[0\ ;\ 1]$
Valeur pour les règles logiques	$\sqrt{1 - e^{-\frac{n_a n_{\bar{b}}}{n}}}$
Valeur pour les règles à l'équilibre	0
Valeur pour les règles à l'indépendance	$\leq \frac{1}{\sqrt{2}}$

TAB. 2.3 – Propriétés de l'intensité d'implication entropique

– Le rejet des règles s'accélère dans une zone d'incertitude autour de l'indépendance $n_{a\bar{b}} = \frac{n_a n_{\bar{b}}}{n}$ (décroissance rapide).

Dans les figures 2.10, les effectifs des données sont multipliés par un coefficient γ à partir d'une configuration initiale. Les indices sont représentés en fonction de γ. Seule l'intensité d'implication est de nature statistique et prend en compte les cardinaux des données de manière absolue : plus le coefficient multiplicateur γ est grand, plus l'écart à l'indépendance observé dans les données est statistiquement significatif, et plus on peut confirmer (figure 2.10.(A)) ou infirmer (figure 2.10.(B)) la bonne qualité de la règle. Cependant, comme le montre la figure 2.11, l'intensité d'implication devient peu discriminante quand les cardinaux étudiés sont grands.



(A) $n_b \geq \frac{n}{2}$ et $n_a \leq \frac{n}{2}$ (B) $n_b < \frac{n}{2}$ ou $n_a > \frac{n}{2}$

FIG. 2.12 – Représentation de l'intensité d'implication entropique selon $n_{a\bar{b}}$

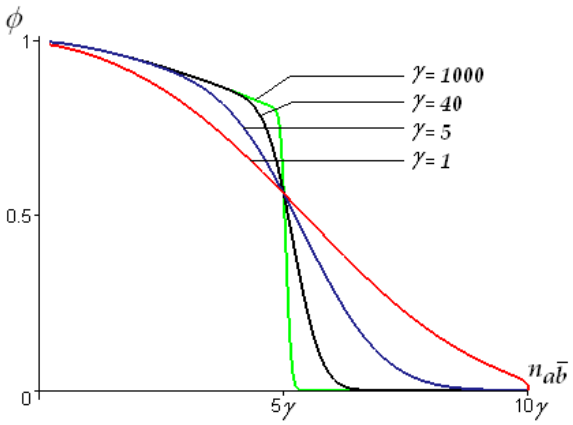


FIG. 2.13 – Représentation de l'intensité d'implication entropique avec la dilatation des effectifs

$(n_a = 20 \times \gamma,\ n_b = 75 \times \gamma,\ n = 100 \times \gamma,$
$n_{a\bar{b}} \in [0 \times \gamma\ ;\ 10 \times \gamma],\ \gamma \in \{1;\ 5;\ 40;\ 1000\})$

Intensité d'implication entropique

L'intensité d'implication entropique prend en compte à la fois l'écart à l'indépendance et l'écart à l'équilibre. Toutefois, selon la classification du chapitre 1, l'intensité d'implication entropique est uniquement un indice d'écart à l'équilibre puisqu'elle ne prend une valeur fixe (en l'occurrence 0) qu'à l'équilibre et non à l'indépendance. Comme nous l'avons vu au chapitre 1 page 25, un indice de règle ne peut pas mesurer à la fois un écart à l'équilibre et un écart à l'indépendance.

Considérons une règle $a \rightarrow b$. En faisant varier $n_{a\bar{b}}$ avec n_a, n_b, et n fixes, on peut distinguer deux comportements différents pour l'intensité d'implication entropique :

- Si $n_b \geq \frac{n}{2}$ et $n_a \leq \frac{n}{2}$, alors l'indépendance est atteinte avant les équilibres de la règle et de sa contraposée quand $n_{a\bar{b}}$ augmente. L'intensité d'implication entropique décroît progressivement avant de s'annuler (figure 2.12.(A)).
- Sinon, l'équilibre de la règle et/ou celui de sa contraposée est atteint avant l'indépendance quand $n_{a\bar{b}}$ augmente. L'intensité d'implication entropique décroît rapidement et s'annule au premier des deux équilibres, c'est-à-dire lorsque $n_{a\bar{b}} = \min(\frac{n_a}{2}, \frac{n_{\bar{b}}}{2})$ (figure 2.12.(B)).

L'intensité d'implication entropique est de nature statistique et varie avec la dilatation des effectifs. Cependant, elle reste discriminante quand les cardinaux étudiés sont grands (voir figure 2.13).

2.3 Le taux informationnel, un indice de règle entropique

Parmi les mesures utilisées en ECD pour évaluer la qualité des règles, les indices issus de la théorie de l'information [SW49] sont particulièrement intelligibles et utiles puisqu'ils peuvent être interprétés en termes d'information. Plus précisément, comme le soulignent Smyth et Goodman [SG91], il existe un parallèle intéressant dans l'utilisation de la théorie de l'information entre les systèmes de communication et l'évaluation des règles. Dans les systèmes de communication, un canal de communication possède une forte capacité s'il peut transmettre une grande quantité d'information de la source au récepteur. Pour une règle, la liaison est de bonne qualité si la prémisse procure beaucoup d'information sur la conclusion. Smyth et Goodman parlent de contenu informationnel (*information content*) d'une règle [SG92].

Les mesures issues de la théorie de l'information traditionnellement utilisées pour évaluer la qualité des règles sont l'entropie conditionnelle de Shannon [CN89], l'information mutuelle moyenne [JS01], le coefficient d'incertitude de Theil [The70] [RZN01] [TKS04], la J-mesure [SG92], et l'indice de Gini [BA99] [JS01]. Ces mesures entropiques ne sont cependant pas des indices de règle au sens de la définition 1.2. Il s'agit en effet de mesures de liaison entre variables multimodales (voir chapitre 1 à la page 11), qui traitent identiquement exemples et contre-exemples. En particulier, elles ne permettent pas de distin-

guer les règles contraires $a \rightarrow b$ et $a \rightarrow \overline{b}$ alors même qu'elles ont des significations opposées. Ces mesures sont davantage adaptées à l'évaluation de règles de classification en apprentissage supervisé, où le modèle recherché doit expliquer toutes les modalités de la variable classe.

Une première démarche pour l'élaboration d'un indice de règle entropique est l'indice d'inclusion de Gras [GKCG01] (présenté à la section 2.2.2 page 49). Il s'agit d'un indice d'écart à l'équilibre. Ici, nous proposons d'utiliser l'entropie pour développer une mesure d'information qui est un indice d'écart à l'indépendance. Ce nouvel indice, appelé *taux informationnel*, a la particularité unique de permettre à la fois le rejet de l'équilibre et le rejet de l'indépendance. Après un rappel des travaux antérieurs sur l'évaluation des règles par la théorie de l'information, nous présentons le taux informationnel à la section 2.3.2, puis étudions ses propriétés. Le nouvel indice est comparé à d'autres mesures en section 2.3.4.

2.3.1 Mesures entropiques pour l'évaluation des règles

Rappels sur l'entropie

Notons V une variable aléatoire discrète à valeurs dans un ensemble D. Dans sa théorie de l'information [SW49], Shannon formalise la notion d'information de la manière suivante :

la quantité d'information de l'événement $V = v$ où $v \in D$ est

$$I(V = v) = -\log_2 \mathrm{P}(V = v) \quad \text{(mesurée en bits)}$$

Objectivement, plus un événement est incertain et plus sa réalisation apporte de l'information (ou vu sous un autre angle : plus un événement est incertain et plus sa prédiction nécessite de l'information). Par suite, l'entropie[3] de la variable V est définie comme la quantité moyenne d'information de V :

$$H(V) = \sum_{v \in D} -\mathrm{P}(V = v).\log_2 \mathrm{P}(V = v)$$

L'entropie est maximale quand toutes les modalités sont équiprobables, et nulle quand l'une des modalités est certaine. $H(V)$ est une mesure de la dispersion de la distribution de probabilité de V.

Bien qu'il existe d'autres mesures d'entropie, nous utilisons dans la suite l'entropie de Shannon H. Pour une variable binaire a, elle s'écrit :

$$H(a) = -\frac{n_a}{n}.\log_2 \frac{n_a}{n} - \frac{n_{\overline{a}}}{n}.\log_2 \frac{n_{\overline{a}}}{n} \quad \text{(voir figure 2.14)}$$

avec l'extension $H(a) = 0$ lorsque $n_a = 0$ ou $n_a = n$. L'entropie conditionnelle d'une variable a sachant l'événement $b = 1$ est définie par :

$$H(a/b = 1) = -\frac{n_{ab}}{n_b}.\log_2 \frac{n_{ab}}{n_b} - \frac{n_{\overline{a}b}}{n_b}.\log_2 \frac{n_{\overline{a}b}}{n_b}$$

[3]Cette mesure est nommée ainsi car elle est identique à l'entropie d'un système en thermodynamique à une constante multiplicative près (les états du système correspondant aux réalisations de la variable). Ce changement d'échelle revient à changer la base du logarithme, et donc à utiliser une autre unité d'information que le bit.

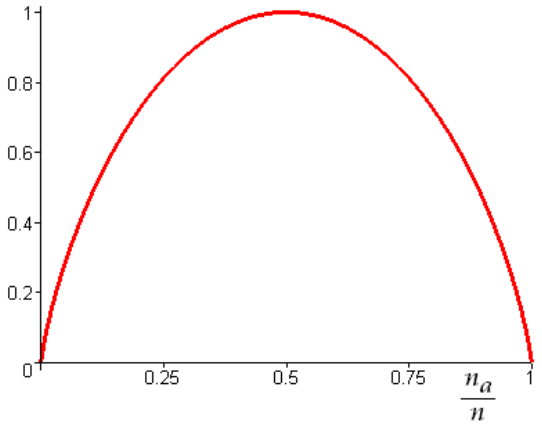


FIG. 2.14 – Représentation de l'entropie de Shannon $H(a)$

Remarquez que les fonctions entropiques associent des variables et des réalisations de variables. Pour plus de clarté, les réalisations d'une variable booléenne b sont notées $b = 1$ et $b = 0$ dans les fonctions entropiques, et non pas b et $\overline{b}$ comme dans les autres notations.

Principales mesures

Les mesures issues de la théorie de l'information traditionnellement utilisées pour évaluer la qualité des règles sont l'entropie conditionnelle de Shannon (utilisée dans l'algorithme CN2 [CN89]), l'information mutuelle moyenne [JS01], le coefficient d'incertitude de Theil [The70] [RZN01] [TKS04], la J-mesure [SG92], l'indice de Gini [BA99] [JS01], et l'indice d'inclusion [GKCG01]. Pour une règle $a \rightarrow b$, l'entropie conditionnelle mesure l'information moyenne de la variable b sachant que a est vrai :

$$H(b/a=1) = -\frac{n_{ab}}{n_a}.\log_2\frac{n_{ab}}{n_a} - \frac{n_{a\overline{b}}}{n_a}.\log_2\frac{n_{a\overline{b}}}{n_a}$$

L'information mutuelle moyenne (diminution d'entropie de Shannon, aussi appelée gain d'entropie ou simplement information mutuelle) mesure l'information moyenne partagée par les variables a et b :

$$\begin{aligned} IM(a,b) &= H(b) - \mathrm{P}(a).H(b/a=1) - \mathrm{P}(\overline{a}).H(b/a=0) \\ &= H(a) - \mathrm{P}(b).H(a/b=1) - \mathrm{P}(\overline{b}).H(a/b=0) \\ &= \frac{n_{ab}}{n}.\log_2\frac{n.n_{ab}}{n_a n_b} + \frac{n_{a\overline{b}}}{n}.\log_2\frac{n.n_{a\overline{b}}}{n_a n_{\overline{b}}} + \frac{n_{\overline{a}b}}{n}.\log_2\frac{n.n_{\overline{a}b}}{n_{\overline{a}} n_b} + \frac{n_{\overline{a}\overline{b}}}{n}.\log_2\frac{n.n_{\overline{a}\overline{b}}}{n_{\overline{a}} n_{\overline{b}}} \end{aligned}$$

Le coefficient d'incertitude u de Theil mesure le taux de réduction d'entropie sur b due à a :

$$u(a,b) = \frac{IM(a,b)}{H(b)}$$

La J-mesure est la part de l'information mutuelle moyenne relative aux événements (a et b) et (a et $\overline{b}$) :

$$J(a,b) = \frac{n_{ab}}{n}.\log_2 \frac{n.n_{ab}}{n_a n_b} + \frac{n_{a\overline{b}}}{n}.\log_2 \frac{n.n_{a\overline{b}}}{n_a n_{\overline{b}}}$$

L'indice de Gini est l'information mutuelle moyenne calculée avec l'entropie quadratique :

$$G(a,b) = \frac{n_a}{n}\left((\frac{n_{ab}}{n_a})^2 + (\frac{n_{a\overline{b}}}{n_a})^2\right) + \frac{n_{\overline{a}}}{n}\left((\frac{n_{\overline{a}b}}{n_{\overline{a}}})^2 + (\frac{n_{\overline{a}\overline{b}}}{n_{\overline{a}}})^2\right) - \frac{{n_b}^2}{n} - \frac{{n_{\overline{b}}}^2}{n}$$

L'indice d'inclusion est fondé sur les entropies conditionnelles $H(b/a=1)$ et $H(a/b=0)$ (voir section 2.2.2).

La quantité d'information que $a=\alpha$ donne sur b

Considérons la quantité d'information donnée par un événement $a=\alpha$ sur une variable b ($\alpha \in \{0;1\}$). Nous notons $M(a=\alpha,\ b)$ les mesures de cette quantité d'information. Blachman [Bla68] a étudié les $M(a=\alpha,\ b)$ dont l'espérance mathématique est l'information mutuelle moyenne entre les variables a et b :

$$IM(a,b) = \mathrm{E}_\alpha\{M(a=\alpha,\ b)\} \tag{2.1}$$

Les deux mesures les plus utilisées sont les suivantes (voir figure 2.15) :

$$j(a=\alpha,\ b) = \mathrm{P}(b/a=\alpha).\log_2 \frac{\mathrm{P}(b/a=\alpha)}{\mathrm{P}(b)} + \mathrm{P}(\overline{b}/a=\alpha).\log_2 \frac{\mathrm{P}(\overline{b}/a=\alpha)}{\mathrm{P}(\overline{b})}$$

$$i(a=\alpha,\ b) = H(b) - H(b/a=\alpha)$$

Blachman montre que j est unique en tant que mesure non-négative qui vérifie l'égalité 2.1, tandis que i est unique en tant que mesure anti-symétrique[4] qui vérifie l'égalité 2.1. Dans [BGGB04], nous avions introduit la mesure i sous le nom de *gain informationnel*.

La mesure j est l'entropie croisée entre les distributions a priori et a posteriori de b. Elle est traditionnellement admise comme "la" mesure de la quantité d'information que $a=\alpha$ donne sur b. En particulier, la J-mesure [SG92] est directement issue de j : $J = j \times \mathrm{P}(a=\alpha)$. Bien que la mesure i ait une interprétation plus aisée (il s'agit de la diminution d'entropie sur b due à l'événement $a=\alpha$), on lui préfère j car i peut s'annuler en dehors de l'indépendance. Ce comportement est dû au caractère symétrique de l'entropie H (invariance par négation).

2.3.2 Taux informationnel

Entropie réduite

Afin de supprimer la symétrie introduite par l'entropie dans la mesure i, nous proposons d'utiliser une fonction entropique orientée $\widehat{H}$ appelée *entropie réduite* (voir figure 2.16) [BGGB04].

[4] i est anti-symétrique vis-à-vis des distributions de la variable b a priori $P = \{\mathrm{P}(b)\}$ et a posteriori $Q = \{\mathrm{P}(b/a=\alpha)\}$: $i(P,Q) = -i(Q,P)$

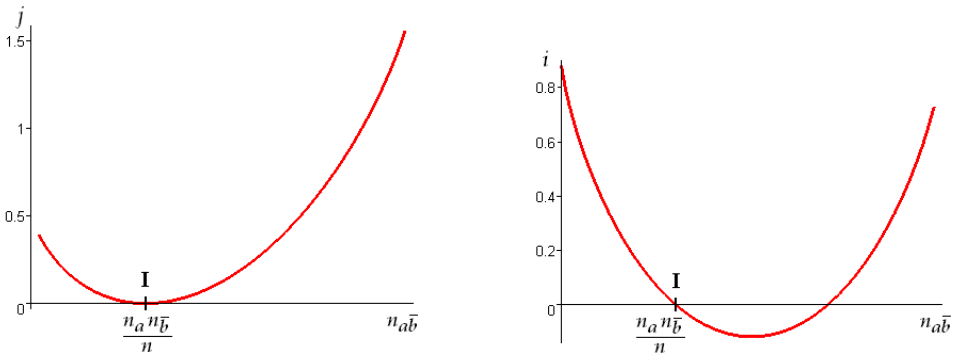


FIG. 2.15 – Représentations des mesures j et i
(I : indépendance)

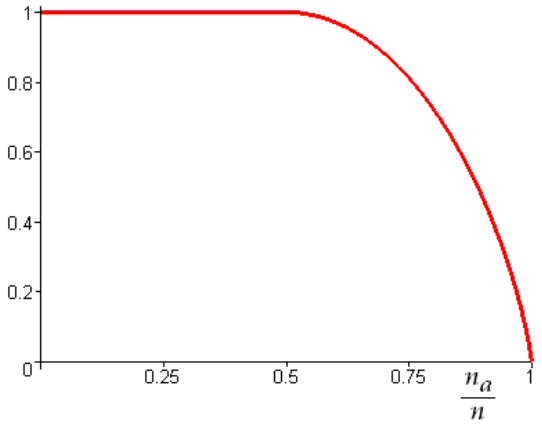


FIG. 2.16 – Représentation de l'entropie réduite $\widehat{H}(a)$

Définition 2.5 L'**entropie réduite** $\widehat{H}(a)$ d'une variable binaire a est définie par :
- si $n_a \leq \frac{n}{2}$ alors $\widehat{H}(a) = 1$,
- si $n_a \geq \frac{n}{2}$ alors $\widehat{H}(a) = H(a)$.

On définit similairement l'entropie réduite conditionnelle d'une variable binaire b sachant la réalisation de a :
- si $n_{ab} \leq \frac{n_a}{2}$ alors $\widehat{H}(b/a = 1) = 1$,
- si $n_{ab} \geq \frac{n_a}{2}$ alors $\widehat{H}(b/a = 1) = H(b/a = 1)$.

L'entropie $H(a)$ d'une variable a peut s'écrire comme la somme de deux entropies réduites :

$$H(a) = \widehat{H}(a) + \widehat{H}(\overline{a}) - 1$$

Contrairement à H, $\widehat{H}$ est une mesure asymétrique qui évalue différemment un déséquilibre en faveur de a et un déséquilibre en faveur de $\overline{a}$: $\widehat{H}(a) \neq \widehat{H}(\overline{a})$. Plus précisément, si a est plus fréquent que $\overline{a}$ ($\frac{n_a}{n} \geq \frac{n_{\overline{a}}}{n}$), alors :
- l'entropie réduite $\widehat{H}(a)$ mesure l'entropie de a : $\widehat{H}(a) = H(a)$;
- l'entropie réduite $\widehat{H}(\overline{a})$ vaut 1.

Si a est moins fréquent que $\overline{a}$ ($\frac{n_a}{n} \leq \frac{n_{\overline{a}}}{n}$), alors les rôles sont inversés. En d'autres termes, $\widehat{H}$ mesure une "incertitude orientée" en faveur d'une des modalités, au sens où si cette modalité n'est pas la plus probable alors l'incertitude est considérée comme maximale.

Indice *TI*

En introduisant l'entropie réduite $\widehat{H}$ dans la mesure i, nous obtenons :

$$i(a = 1,\ b) = \widehat{H}(b) + \widehat{H}(\overline{b}) - \widehat{H}(b/a = 1) - \widehat{H}(\overline{b}/a = 1)$$

D'où :

$$i(a = 1,\ b) \ = \ \widehat{i}(a = 1,\ b) \ + \ \widehat{i}(a = 1,\ \overline{b})$$

$$\text{avec } \widehat{i}(a = 1,\ b) = \widehat{H}(b) - \widehat{H}(b/a = 1)$$

L'indice i qui mesure une diminution d'entropie H est donc la somme de deux diminutions d'entropie réduite $\widehat{H}$:
- $\widehat{i}(a = 1,\ b)$ qui est la diminution d'entropie réduite sur b due à $a = 1$,
- $\widehat{i}(a = 1,\ \overline{b})$ qui est la diminution d'entropie réduite sur $\overline{b}$ due à $a = 1$.

Contrairement aux mesures i et j, $\widehat{i}$ a l'avantage d'être un indice de règle au sens de la définition 1.2 :

$$\widehat{i}(a = 1,\ b) = \widehat{i}(a \rightarrow b) \quad \text{(voir figure 2.17)}$$

Plus $\widehat{i}(a \rightarrow b)$ est élevé, plus l'événement $a = 1$ apporte d'information en faveur de $b = 1$ et plus la qualité de la règle est garantie. Si $\widehat{i}(a \rightarrow b)$ est négatif, cela signifie que l'événement $a = 1$ n'apporte aucune information en faveur de $b = 1$, et même qu'il en "retire". En d'autres termes, l'incertitude est moindre à prédire $b = 1$ au hasard qu'à prédire $b = 1$ en utilisant la règle. Selon nous, $\widehat{i}$ est la mesure de ce que Smyth et Goodman appellent le contenu informationnel des règles [SG92].

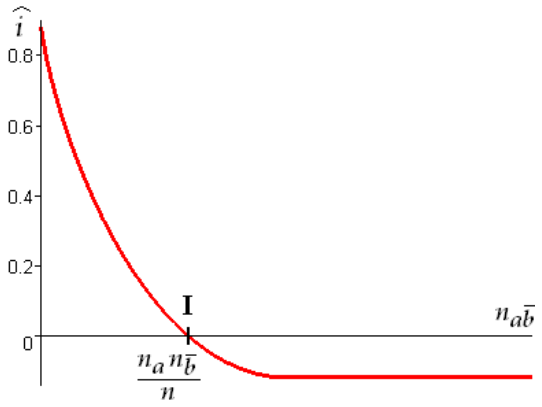


FIG. 2.17 – Représentation de l'indice de règle $\widehat{i}$ en fonction de $n_{a\overline{b}}$

Par ailleurs, à l'instar de la contribution orientée au χ^2 de Lerman [Ler81] (voir chapitre 1 page 37), $\widehat{i}$ permet de distribuer l'information mutuelle moyenne sur les règles qui résident entre les deux variables :

$$IM(a,b) = \frac{n_a}{n}.\widehat{i}(a \rightarrow b) + \frac{n_a}{n}.\widehat{i}(a \rightarrow \overline{b}) + \frac{n_{\overline{a}}}{n}.\widehat{i}(\overline{a} \rightarrow b) + \frac{n_{\overline{a}}}{n}.\widehat{i}(\overline{a} \rightarrow \overline{b})$$

$\frac{n_a}{n}.\widehat{i}(a \rightarrow b)$ est la contribution orientée de la règle $a \rightarrow b$ à l'information mutuelle moyenne. Chaque règle participe à l'information mutuelle moyenne en apportant ou retranchant sa part d'information. Comme le χ^2, l'information mutuelle moyenne peut aussi s'écrire avec les contributions des quatre règles contraires.

Pour ces différentes caractéristiques, nous proposons de retenir l'indice $\widehat{i}$ pour mesurer la qualité des règles. Cependant, $\widehat{i}$ a le désavantage d'avoir une valeur maximale qui n'est pas fixe mais dépend de n_b, ce qui rend difficile la comparaison de règles avec des conclusions différentes. Pour faciliter le filtrage des règles les plus informatives, nous normalisons $\widehat{i}$ en attribuant le score maximal 1 aux meilleures règles. Ceci revient à calculer le taux de réduction de l'entropie réduite $\widehat{H}$ [BGGB04].

Définition 2.6 Le **taux informationnel**[5] **(*TI*)** d'une règle $a \rightarrow b$ est défini par :

$$TI(a \rightarrow b) = \frac{\widehat{H}(b) - \widehat{H}(b/a=1)}{\widehat{H}(b)} \quad \text{si } n_{\overline{b}} \neq 0$$

La mesure n'est pas définie si $n_b = n$, mais ces règles sont évidemment à rejeter ($\widehat{i}$ est d'ailleurs nul pour ces règles). Une règle est dite *informative* si son taux informationnel est strictement positif.

[5] *Directed Information Ratio* (*DIR*) dans [BGGB05]

Objet	écart à l'indépendance
Portée	règle au sens strict
Nature	descriptive
Domaine de variation	$]-\infty\ ;\ 1]$
Valeur pour les règles logiques	1
Valeur pour les règles à l'équilibre	$1-\widehat{H}(b)^{-1} \leq 0$
Valeur pour les règles à l'indépendance	0

TAB. 2.4 – Propriétés de TI

Prise en compte de la contraposée dans l'indice *TIC*

En associant les taux informationnels d'une règle et de sa contraposée au sein d'une mesure synthétique, l'indice de règle TI peut être dérivé en un indice de quasi-implication. Afin que le contenu informationnel de la quasi-implication soit nul dès que la règle ou sa contraposée n'est pas informative, nous avons choisi d'utiliser la moyenne géométrique pour combiner les deux taux informationnels en rejetant tous les taux négatifs.

Définition 2.7 Le **taux informationnel modulé par la contraposée (*TIC*)** d'une règle $a \rightarrow b$ est défini par :

$$TIC(a \rightarrow b) = \sqrt{TI(a \rightarrow b) \times TI(\overline{b} \rightarrow \overline{a})} \quad si\ \ TI(a \rightarrow b) \geq 0\ \ et\ \ TI(\overline{b} \rightarrow \overline{a}) \geq 0$$

$$TIC(a \rightarrow b) = 0 \quad sinon$$

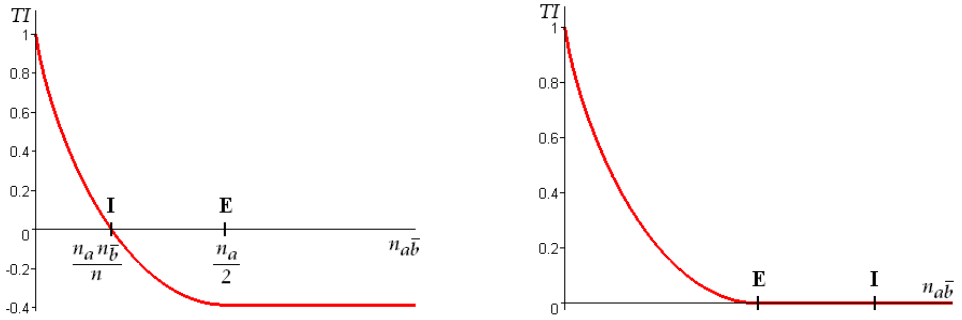
2.3.3 Propriétés

Taux informationnel *TI*

Les principales propriétés de TI sont données dans le tableau 2.4. TI est une fonction décroissante convexe du nombre de contre-exemples. Il fait partie des indices de règle "exigeants" qui diminuent rapidement dès les premiers contre-exemples et permettent ainsi de mieux hiérarchiser les bonnes règles (plus grande dispersion des valeurs). Comme les mesures entropiques, TI est de nature descriptive.

Considérons une règle $a \rightarrow b$. En faisant varier $n_{a\overline{b}}$ avec n_a, n_b, et n fixes, on peut distinguer deux comportements différents pour TI [BGGB04] :

- Si $n_b \geq \frac{n}{2}$, alors l'indépendance est atteinte avant l'équilibre quand $n_{a\overline{b}}$ augmente. Le taux informationnel s'annule à l'indépendance puis admet des valeurs négatives (figure 2.18.(A)).
- Si $n_b \leq \frac{n}{2}$, alors l'équilibre est atteint avant l'indépendance quand $n_{a\overline{b}}$ augmente. Le taux informationnel s'annule mais n'admet pas de valeurs négatives (figure 2.18.(B)).



(A) $n_b \geq \frac{n}{2}$ (B) $n_b \leq \frac{n}{2}$

FIG. 2.18 – Représentation de TI en fonction de $n_{a\bar{b}}$

Au sens des définitions 1.4 et 1.5 du chapitre 1, TI est un indice d'écart à l'indépendance et non d'écart à l'équilibre. Cependant, TI permet de repérer à la fois les situations d'indépendance et d'équilibre des règles. En effet, dans ces situations, TI prend des valeurs négatives ou nulles (voir tableau 2.4). En ne retenant que les taux informationnels strictement positifs (règles informatives), l'utilisateur rejette donc toutes les règles dont l'écart à l'indépendance est mauvais (règles entre variables corrélées négativement), mais aussi toutes les règles dont l'écart à l'équilibre est mauvais (règles qui possèdent plus de contre-exemples que d'exemples). La mesure doit donc être utilisée avec un seuil strictement positif pour filtrer les règles. A notre connaissance, TI est le seul indice de règle qui puisse rejeter à la fois indépendance et équilibre avec un seuil fixe. C'est une approche tout à fait originale pour l'évaluation de la qualité des règles.

Exemple. Nous reprenons l'exemple de la section 1.4.1 page 25, qui porte sur les règles $r_1 = (800,\ 1000,\ 4500,\ 5000)$ et $r_2 = (400,\ 1000,\ 1000,\ 5000)$. Le lift (indice d'écart à l'indépendance qui vaut 1 à l'indépendance) permet de rejeter r_1 mais pas r_2 :

$$lift(r_1) = 0.9 < 1 \quad \text{et} \quad lift(r_2) = 2 > 1$$

En revanche, la confiance (indice d'écart à l'équilibre qui vaut 0.5 à l'équilibre) permet de rejeter r_2 mais pas r_1 :

$$confiance(r_1) = 0.8 > 0.5 \quad \text{et} \quad confiance(r_2) = 0.4 < 0.5$$

Le taux informationnel permet quant à lui de rejeter les deux règles :

$$TI(r_1) \simeq -0.5 \leq 0 \quad \text{et} \quad TI(r_2) = 0 \qquad \square$$

Taux informationnel modulé par la contraposé *TIC*

Les principales propriétés de TIC sont données dans le tableau 2.5. C'est aussi une mesure qui décroît rapidement dès les premiers contre-exemples. Le

Objet	écart à l'indépendance
Portée	quasi-implication
Nature	descriptive
Domaine de variation	[0 ; 1]
Valeur pour les règles logiques	1
Valeur pour les règles à l'équilibre	0
Valeur pour les règles à l'indépendance	0

TAB. 2.5 – Propriétés de *TIC*

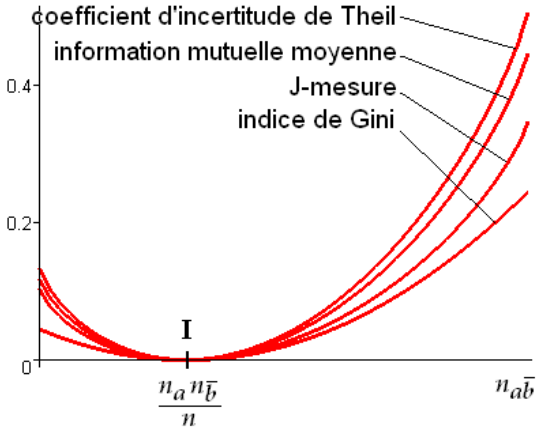


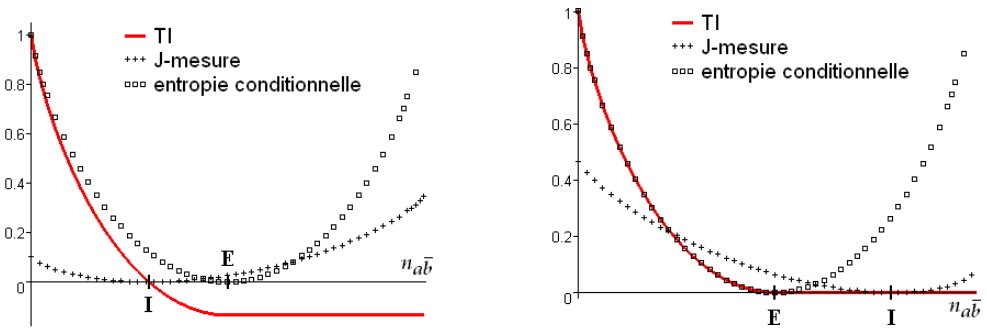
FIG. 2.19 – Les mesures entropiques utilisées pour évaluer des règles

taux informationnel TIC d'une quasi-implication est nul dès qu'une des deux règles qui la constitue n'est pas informative. Ainsi, TIC permet de repérer à la fois les situations d'équilibre pour la règle ou sa contraposée et les situations d'indépendance.

2.3.4 Comparaisons à d'autres mesures

Comparaisons formelles

Dans cette section, nous comparons le taux informationnel TI aux mesures issues de la théorie de l'information traditionnellement utilisées pour évaluer la qualité des règles : l'entropie conditionnelle de Shannon, l'information mutuelle moyenne, le coefficient d'incertitude de Theil, la J-mesure, et l'indice de Gini. Etant donné que les quatre dernières mesures sont très similaires (voir figure 2.19), nous n'en considérons qu'une seule parmi les quatre dans les comparaisons qui suivent. Nous choisissons la J-mesure, qui est la plus utilisée dans le contexte des règles d'association. En ce qui concerne l'entropie conditionnelle, pour une



(A) $n_b \geq \frac{n}{2}$ (B) $n_b \leq \frac{n}{2}$

FIG. 2.20 – Représentation de TI, la J-mesure, et l'entropie conditionnelle en fonction de $n_{a\bar{b}}$

	Nombre d'items	**Nombre d'individus**	**Nombre de règles découvertes**
T10.I4.D5k	12	5000	97688
T10.I4.D100k	1000	100000	478894
PANNES	92	2883	43930
PROFILS	30	2299	28938

TAB. 2.6 – Caractéristiques des données

règle $a \rightarrow b$ ce n'est pas la fonction $H(b/a = 1)$ décrite à la section 2.3.1 qui est représentée dans les comparaisons, mais la fonction complémentaire $1 - H(b/a = 1)$. En effet, contrairement aux autres mesures, $H(b/a = 1)$ attribue ses plus petites valeurs aux meilleures règles (pour générer des règles de qualité, l'algorithme CN2 cherche à minimiser $H(b/a = 1)$ [CN89]).

Les figures 2.20.(A) et 2.20.(B) comparent TI à l'entropie conditionnelle et à la J-mesure quand le nombre de contre-exemples $n_{a\bar{b}}$ augmente. Les figures illustrent clairement que l'entropie conditionnelle et la J-mesure ne sont pas des indices de règle, puisqu'elles peuvent croître quand les contre-exemples augmentent. De plus, la J-mesure repère l'indépendance (elle s'y annule) mais pas l'équilibre (elle peut même prendre des valeurs élevées à l'équilibre), alors que l'entropie conditionnelle repère l'équilibre (elle s'y annule) mais pas l'indépendance (elle peut même prendre des valeurs élevées à l'indépendance). Dans tous les cas, filtrer les règles sur TI avec un seuil strictement positif suffit pour rejeter à la fois équilibre et indépendance. Comme l'illustre la figure 2.20.(B), TI est analogue à l'entropie conditionnelle quand $n_b \leq \frac{n}{2}$ (les fonctions sont partiellement identiques). C'est ce qui permet à TI de s'annuler à l'équilibre quand $n_b \leq \frac{n}{2}$.

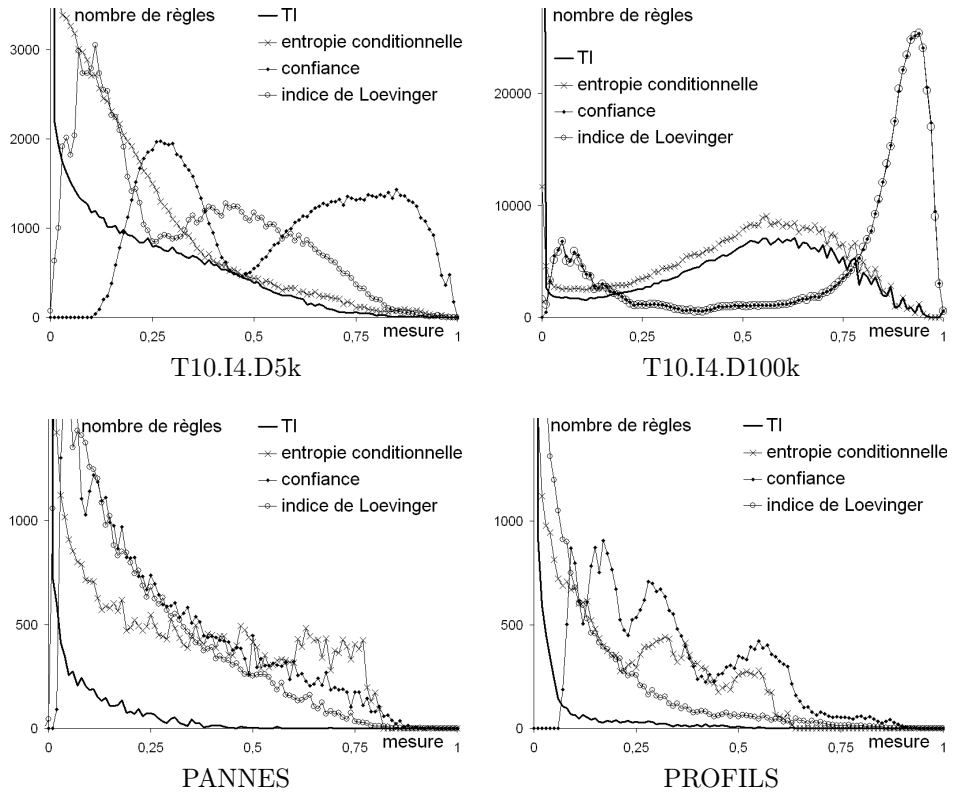


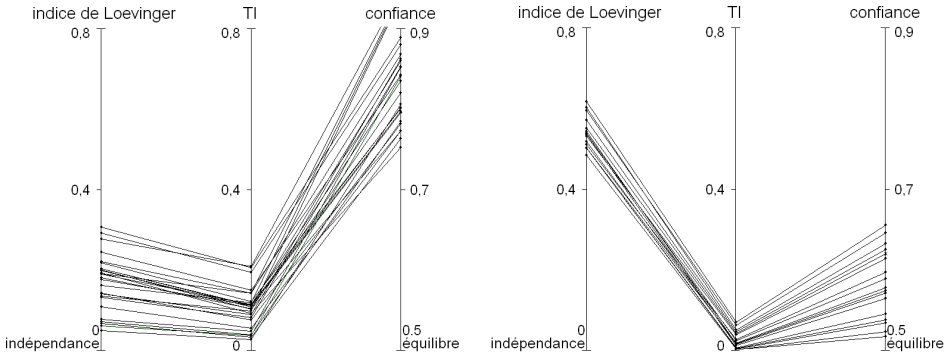
FIG. 2.21 – Distributions des mesures sur les ensembles de règles

Comparaisons expérimentales

Nous comparons les distributions de TI aux distributions d'autres mesures sur un ensemble de règles d'association extraites à partir de quatre jeux de données (décrits dans le tableau 2.6). Les deux premiers jeux de données ont été créés à l'aide du générateur[6] de données synthétiques d'IBM décrit dans [AS94a], qui simule des achats dans un supermarché. Les deux autres jeux de données sont une base de données de pannes d'ascenseurs fournie par une société de maintenance, et une base de profils psychologiques utilisée en gestion des ressources humaines, appartenant à la société *PerformanSe SA*[7]. Les règles ont été extraites à l'aide de l'algorithme *Apriori* [AS94a] avec un seuil de support faible pour éviter l'élimination prématurée de règles potentiellement intéressantes (voir chapitre 3 pour plus de détails sur l'algorithme *Apriori*).

Puisque nous souhaitons ici comparer les distributions des mesures, nous choisissons des mesures qui, comme TI, ont 1 pour valeur maximale. Parmi les mesures entropiques, seule l'entropie conditionnelle satisfait à cette condition. Nous ajoutons donc à nos comparaisons deux indices de règle qui vérifient cette condition : la confiance et l'indice de Loevinger (voir définitions dans le tableau

[6] http ://www.almaden.ibm.com/software/quest/Resources/index.shtml
[7] www.performanse.fr



(A) échantillon de règles extrait de T10.I4.D5k

(B) échantillon de règles extrait de PANNES

FIG. 2.22 – Deux échantillons de règles représentés en coordonnées parallèles

1.3 page 16). Ils mesurent respectivement un écart à l'équilibre et un écart à l'indépendance. Comme le montre la figure 2.21, le taux informationnel TI est l'indice le plus filtrant : pour les quatre jeux de données, quel que soit le seuil choisi entre 0 et 1, TI élimine plus de règles que les autres mesures. Ceci est particulièrement utile pour le post-traitement de grands ensembles de règles.

Expliquons pourquoi TI est un indice très filtrant. Dans les figures 2.22 en coordonnées parallèles, chaque ligne brisée représente une règle. La figure 2.22.(A) montre des règles représentatives de T10.I4.D5k qui sont jugées bonnes par la confiance mais pas par l'indice de Loevinger, alors que la figure 2.22.(B) exhibe des règles de PANNES qui sont jugées bonnes par l'indice de Loevinger mais pas par la confiance. En prenant en compte équilibre et indépendance, TI donne de mauvaises valeurs à toutes ces règles.

2.4 Conclusion

Nous avons présenté dans ce chapitre trois nouveaux indices de règle aux propriétés originales.

- L'indice probabiliste d'écart à l'équilibre IPEE est l'unique indice d'écart à l'équilibre qui soit de nature statistique. Fondé sur un modèle probabiliste, il évalue la significativité statistique de l'écart à l'équilibre.
- L'intensité d'implication entropique est une version de l'intensité d'implication corrigée pour rester discriminante sur les grands jeux de données. Elle tient compte à la fois de l'écart à l'équilibre et de l'écart à l'indépendance.
- Le taux informationnel est un indice de règle fondé sur la théorie de l'information. Il permet de rejeter à la fois les règles dont l'écart à l'équilibre est mauvais (règles qui possèdent plus de contre-exemples que d'exemples) et les règles dont l'écart à l'indépendance est mauvais (règles entre variables corrélées négativement). Des comparaisons expérimentales montrent que le taux informationnel est une mesure très filtrante, ce qui est utile pour

le post-traitement de grands ensembles de règles.

Ces trois indices ne sont pas concurrents : alors que le taux informationnel est un indice descriptif (comme toutes les mesures entropiques), IPEE et l'intensité d'implication employés conjointement permettent de réaliser une évaluation statistique complète des règles. IPEE peut être vu comme l'analogue de l'intensité d'implication pour l'écart à l'équilibre.

D'un point de vue plus pratique, le taux informationnel et l'intensité d'implication entropique sont des mesures perfectionnées qui combinent de multiples propriétés. Elles sont tout à fait adaptées aux cas où l'utilisateur ne souhaite employer qu'un seul indice de règle synthétique qui puisse évaluer simultanément plusieurs aspects de la qualité des règles. Bien qu'elles combinent moins de propriétés que le taux informationnel ou l'intensité d'implication entropique, des mesures comme l'intensité d'implication ou IPEE présentent en revanche l'avantage d'être plus intelligibles. Elles conviennent plus aux post-traitements de règles où divers indices sont utilisés conjointement. Dans un tel cas, il est en effet préférable d'employer des indices intelligibles qui mesurent des aspects "orthogonaux"[8] de la qualité des règles, plutôt que des indices perfectionnés mais moins intelligibles. L'outil de visualisation de règles présenté en partie 3 exploitant plusieurs indices, ceci explique pourquoi, parmi les trois mesures proposées dans ce chapitre, nous avons choisi de n'intégrer que IPEE à cet outil (accompagné entre autres de l'intensité d'implication).

[8] Voir chapitre 1 à la page 38 pour un exemple.

Deuxième partie

Extraction et post-traitement des règles d'association

Extraction des règles d'association

3

Sommaire

En ECD, l'une des principales techniques produisant des connaissances sous forme de règles est *l'extraction de règles d'association*, introduite par Agrawal, Imielienski et Swami [AIS93]. Les algorithmes d'extraction de règles d'association exécutent tous la même tâche déterministe : étant donnés un seuil de support minimal et un seuil de confiance minimale, produire l'ensemble exhaustif de toutes les règles qui possèdent un support supérieur au seuil (contrainte de généralité) et une confiance supérieure au seuil (contrainte de validité). De nombreuses généralisations et adaptations des règles d'association ont aussi été étudiées, dont les principales sont les règles d'association numériques (portant sur des variables quantitatives) [SA96a] [FMMT01], les règles d'association généralisées (exploitant une hiérarchie de concepts) [SA95] [HF95], et les motifs séquentiels extraits à partir de données temporelles [SA96b] [MTV97] [Zak01].

Depuis l'algorithme de référence d'Agrawal et Srikant [AS94a], nommé *Apriori*[1], de nombreux algorithmes ont été proposés pour extraire efficacement des règles d'association. En parallèle, des méthodes ont été développées pour extraire des règles "sous contraintes", c'est-à-dire avec d'autres contraintes que

[1] L'année où *Apriori* fut publié, un algorithme analogue a été proposé par Mannila, Toivonen, et Verkamo [MTV94]. Les deux approches ont été réunies dans [AMS+96].

celles du support et de la confiance. Ce chapitre passe en revue les deux types d'algorithmes d'extraction de règles d'association.

3.1 Terminologie et notations

Nous reprenons le vocabulaire et les notations de la partie précédente :
- E est un ensemble de n individus décrits par un ensemble $I = \{i_1, i_2, ..., i_p\}$ de variables booléennes appelées *items*. E est stocké dans une table dans une base de données relationnelle (de par sa taille, E ne tient pas en mémoire centrale).
- Une conjonction d'items est appelée un *itemset*, comme par exemple $(i_1 \wedge i_4 \wedge i_5)$. Il s'agit également d'une variable booléenne.
- Etant donnés deux itemsets a et b, nous notons n_a le nombre d'individus qui vérifient a, et n_{ab} le nombre d'individus qui vérifient à la fois a et b.

De plus, nous appelons *k-itemset* un itemset de longueur k, c'est-à-dire un itemset qui contient k items. Etant donné un itemset a, l'itemset b est un sur-itemset de a si tous les items de a sont présents dans b. a est alors un sous-itemset de b.

3.2 Règles d'association

Définition 3.1 (d'après [AS94a]) Une **règle d'association** est un couple de variables booléennes (a, b) noté $a \rightarrow b$ où a et b sont des itemsets qui n'ont pas d'item en commun.

Les règles d'association sont un cas particulier des règles introduites au chapitre 1 (définition 1.1 page 12) pour deux raisons.
- Tout d'abord, les règles d'association ne s'établissent qu'entre des itemsets et jamais entre des négations d'itemsets. Elles ne contiennent donc jamais de disjonction, mais uniquement des conjonctions. Quelques extensions des règles d'association aux disjonctions ont toutefois été proposées (voir par exemple [NCJK01]).
- Ensuite, les algorithmes d'extraction de règles d'association ne prennent en compte que la vérité des items, et jamais leur fausseté, qui est considérée comme moins intéressante[2]. Ainsi, les règles d'association ne peuvent pas contenir de négations d'items. Dans la pratique il s'agit rarement d'une restriction, car si la fausseté d'un item particulier se révèle intéressante pour une application, il suffit :
  - soit d'intervertir dans les données l'item et sa négation,
  - soit, si l'on souhaite conserver à la fois l'information vérité et l'information fausseté, de rajouter parmi les variables qui décrivent les données

[2]La première application des règles d'association a été l'analyse du panier de la ménagère, qui consiste à découvrir des combinaisons d'articles qui sont souvent achetés ensemble dans un supermarché. Dans cette application, la fausseté des items (un article n'est pas acheté par un client) est beaucoup trop fréquente devant la vérité (un article est acheté par un client), ce qui explique qu'elle ne soit pas prise en compte.

la négation de l'item.
La prise en compte systématique de la négation pour tous les items dépasse le cadre classique de l'extraction de règles d'association, mais a fait l'objet de différents travaux [SON98] [BBJ00] [WZZ04].

L'extraction de règles d'association est fondée sur deux mesures, le support et la confiance [AIS93]. Nous rappelons ci-dessous leurs définitions.

- Le support évalue la généralité d'une règle. Il s'agit de la proportion d'individus qui vérifient la règle dans le jeu de données :

$$support(a \rightarrow b) = \frac{n_{ab}}{n}$$

- La confiance évalue la validité d'une règle. Il s'agit de la proportion d'individus qui vérifient la conclusion parmi ceux qui vérifient la prémisse :

$$confiance(a \rightarrow b) = \frac{n_{ab}}{n_a}$$

La mesure du support s'applique aussi aux itemsets. Pour un itemset a :

$$support(a) = \frac{n_a}{n}$$

Etant donné un seuil de support minimal σ_{sp}, un itemset a est dit *fréquent* si $support(a) \geq \sigma_{sp}$. Etant donné un seuil de confiance minimale σ_{cf}, une règle $a \rightarrow b$ est dite *valide* si $confiance(a \rightarrow b) \geq \sigma_{cf}$.

3.3 Algorithmes exhaustifs

Soient σ_{sp} et σ_{cf} les seuils de support et de confiance fixés par l'utilisateur. Les algorithmes exhaustifs pour l'extraction de règles d'association se déroulent en deux étapes successives :

1. Trouver tous les itemsets fréquents au seuil σ_{sp}.

2. A partir des itemsets fréquents, générer toutes les règles d'association valides au seuil σ_{cf}.

Les différences de performances entre les algorithmes dépendent principalement de la première étape, qui domine les coûts en temps et en mémoire des algorithmes [NLHP98]. C'est donc sur l'étape 1 que se portent tous les efforts pour l'optimisation des algorithmes d'extraction de règles d'association. L'étape 2, quant à elle, est identique quel que soit l'algorithme considéré.

3.3.1 Algorithme *Apriori*

Extraction des itemsets fréquents

La taille de l'espace de recherche pour l'extraction des itemsets fréquents est exponentielle avec le nombre p d'items (il y a potentiellement $2^p - 1$ item-

sets fréquents). Pour réduire cet espace, l'algorithme *Apriori* [AS94a] tire profit d'une propriété d'anti-monotonie du support (voir figure 3.1) :

si un itemset n'est pas fréquent, aucun de ses sur-itemsets n'est fréquent.

L'extraction des itemsets fréquents est ainsi réalisée niveau par niveau (recherche en largeur d'abord) : pour déterminer l'ensemble F_k des k-itemsets fréquents, l'algorithme construit un ensemble C_k de k-itemsets candidats (c'est-à-dire susceptibles d'être fréquents) à partir de l'ensemble F_{k-1} des (k-1)-itemsets fréquents extraits au niveau précédent (voir algorithme 3.1). Plus précisément, à chaque niveau k, l'algorithme effectue deux opérations :

1. **La génération de C_k à partir de F_{k-1}** (lignes 5 et 6 dans l'algorithme 3.1) : Pour construire les k-itemsets candidats, les (k-1)-itemsets fréquents sont combinés entre eux. A partir de deux (k-1)-itemsets qui possèdent k-2 items en commun, on génère un k-itemset en regroupant tous les items (avec suppression des k-2 doublons). Par exemple, deux 3-itemsets ($i_1 \wedge i_3 \wedge i_4$) et ($i_1 \wedge i_3 \wedge i_7$) permettent de générer le 4-itemset candidat ($i_1 \wedge i_3 \wedge i_4 \wedge i_7$).
2. **L'élagage de C_k pour déterminer F_k** (lignes 7 à 13 dans l'algorithme 3.1) : Dans un premier temps, on élague C_k en réutilisant la propriété d'anti-monotonie. En effet, un k-itemset candidat ne peut être fréquent que si tous ses sous-itemsets de longueur k-1 sont eux-mêmes fréquents. Tous les candidats qui ne respectent pas cette condition sont supprimés de C_k (procédure `élagagePréliminaire()` dans l'algorithme 3.1). Dans un deuxième temps, les données sont lues (un passage sur la table en base de données) afin de déterminer le support des itemsets candidats, par le biais d'un comptage de leurs occurrences (ligne 12). F_k est formé de tous les itemsets candidats qui sont effectivement fréquents (ligne 13).

Au final, l'algorithme fournit l'ensemble des itemsets fréquents mais aussi leurs supports. Ceux-ci sont nécessaires pour calculer par la suite les indices de règle. Pour les cas où l'utilisateur n'est intéressé que par l'ensemble des itemsets fréquents et non par leurs supports, des algorithmes d'extraction des itemsets fréquents maximaux ont été développés [Bay98] [GZ01] [Bur01] [TNHB00]. Les itemsets fréquents maximaux sont les itemsets fréquents dont aucun sur-itemset n'est fréquent. Ils permettent de retrouver facilement les itemsets fréquents, puisque l'ensemble des itemsets fréquents est l'ensemble des itemsets fréquents maximaux et de leurs sous-itemsets.

Dans la pratique, l'efficacité de l'algorithme *Apriori* dépend du seuil de support utilisé et des données étudiées. D'un point de vue qualitatif, on qualifie les données de "creuses" (respectivement "denses") quand la vérité des items est peu fréquente (respectivement très fréquente) en comparaison à la fausseté (proportion des 1 par rapport aux 0 dans la matrice booléenne). Pour un seuil de support donné :

- plus les données sont creuses, plus la propriété d'anti-monotonie est efficace pour réduire l'espace de recherche, et plus l'algorithme peut supporter un nombre élevé d'items décrivant les données ;
- plus les données sont denses, moins la propriété d'anti-monotonie est efficace pour réduire l'espace de recherche, et moins l'algorithme supporte

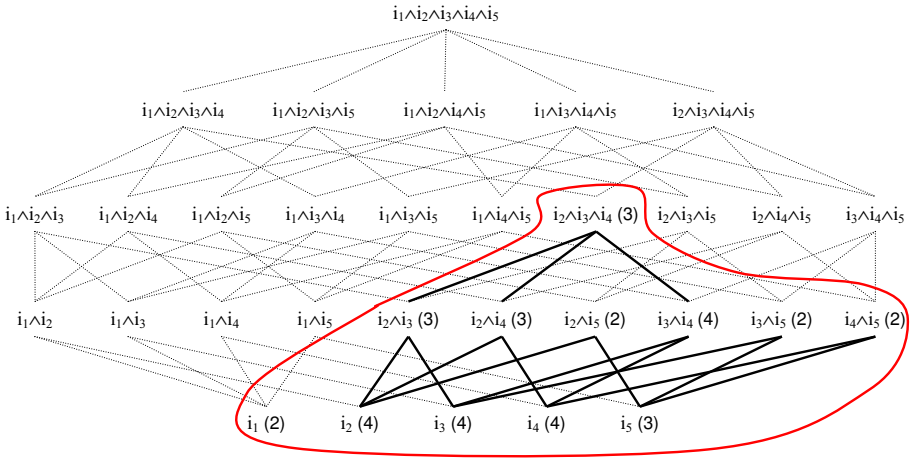
Seuil de support utilisé : $\frac{3}{5}$

i_1	i_2	i_3	i_4	i_5
1	1	1	1	0
1	1	0	0	1
0	1	1	1	1
0	0	1	1	1
0	1	1	1	0

(A)

$F_1 = \{\ i_2\ ,\ i_3\ ,\ i_4\ ,\ i_5\ \}$
$C_2 = \{\ i_2 \wedge i_3\ ,\ i_2 \wedge i_4\ ,\ i_2 \wedge i_5\ ,\ i_3 \wedge i_4\ ,\ i_3 \wedge i_5\ ,\ i_4 \wedge i_5\ \}$
$F_2 = \{\ i_2 \wedge i_3\ ,\ i_2 \wedge i_4\ ,\ i_3 \wedge i_4\}$
$C_3 = \{\ i_2 \wedge i_3 \wedge i_4\ \}$
$F_3 = \{\ i_2 \wedge i_3 \wedge i_4\ \}$
$C_4 = \emptyset$

(B)



(C)

FIG. 3.1 – Réduction de l'espace de recherche pour l'extraction des itemsets fréquents dans l'algorithme *Apriori*
L'algorithme traite les données (A) contenant 5 individus décrits par 5 items. Les ensembles d'itemsets candidats et fréquents pour le seuil de support minimal $\frac{3}{5}$ sont donnés en (B). L'espace de recherche complet pour l'extraction des itemsets fréquents est représenté en (C). Seuls les itemsets entourés en rouge sont effectivement explorés par l'algorithme (leurs occurrences sont comptées dans les données). Le support de ces itemsets (en nombre d'occurrences) est indiqué entre parenthèses.

un nombre élevé d'items dans les données.
Si l'explosion combinatoire du nombre d'itemsets fréquents rend l'algorithme inutilisable, la seule façon de pouvoir traiter les données est d'augmenter le seuil de support.

Entrées : – $\mathcal{BD}$, base de données,
– σ_{sp}, seuil de support minimal
Sorties : – L, ensemble de couples $(i, sp(i))$ où i est un itemset et $sp(i)$ son support

```
1  F_1 = extraireItemsFréquents(BD, σ_sp) ;
2  n = nombreIndividus(BD) ;
3  k = 2 ;
4  tant que F_{k-1} ≠ ∅ faire
5  |  //Génération de C_k à partir de F_{k-1} :
6  |  C_k = générerItemsetsCandidats(F_{k-1}) ;
7  |  //Elagage de C_k pour déterminer F_k :
8  |  C_k = élagagePréliminaire(C_k, F_{k-1}) ;
9  |  pour chaque individu ∈ BD faire
10 |  |  pour chaque i ∈ C_k faire
11 |  |  |  si individu vérifie i alors
12 |  |  |  |  i.compteur++ ;
13 |  |  F_k = {(i, i.compteur/n) | i ∈ C_k et i.compteur/n ≥ σ_sp} ;
14 |  |  k++ ;
15 L = ⋃_k F_k ;
16 retourne L ;
```

Algorithme 3.1: Extraction des itemsets fréquents dans *Apriori*

Génération des règles d'association

A partir d'un itemset fréquent i, l'algorithme construit toutes les règles de la forme $a \rightarrow b$ où a et b sont deux sous-itemsets de i qui ne possèdent pas d'item en commun et qui redonnent i par conjonction : $a \wedge b = i$. La confiance d'une telle règle est calculée de la manière suivante : $confiance\,(a \rightarrow b) = \frac{support(i)}{support(a)}$. Le fait que i soit un itemset fréquent garantit que son sous-itemset a l'est aussi, et donc que son support est connu (déterminé à l'étape d'extraction des itemsets fréquents). Les règles retournées par l'algorithme sont celles dont la confiance est supérieure au seuil de confiance minimale σ_{cf}. Le support de b, qui a aussi été déterminé à l'étape d'extraction des itemsets fréquents, peut être utilisé pour calculer d'autres indices de règle que le support et la confiance (voir partie 1).

Par exemple, considérons l'itemset fréquent $i_2 \wedge i_3 \wedge i_4$ de la figure 3.1. A partir de cet itemset, on peut générer les six règles suivantes :
– $i_2 \wedge i_3 \rightarrow i_4 \quad (confiance = 100\%)$
– $i_2 \wedge i_4 \rightarrow i_3 \quad (confiance = 100\%)$

- $i_3 \wedge i_4 \rightarrow i_2 \quad (confiance = 75\%)$
- $i_2 \rightarrow i_3 \wedge i_4 \quad (confiance = 75\%)$
- $i_3 \rightarrow i_2 \wedge i_4 \quad (confiance = 75\%)$
- $i_4 \rightarrow i_2 \wedge i_3 \quad (confiance = 75\%)$

Avec un seuil de confiance minimale $\sigma_{cf} = 90\%$, seules les deux premières règles sont valides et retournées par l'algorithme.

3.3.2 Autres algorithmes

L'algorithme *DIC* [BMUT97] est une version assouplie de *Apriori* dans laquelle les passages sur les données ne sont pas affectés exclusivement à un seul niveau k de k-itemsets. Le principe de base est le suivant : au cours d'un passage sur les données, si le compteur des occurrences d'un itemset est déjà suffisamment élevé pour que l'on sache que celui-ci est fréquent, alors l'itemset peut d'ores et déjà être utilisé pour générer des candidats du niveau supérieur. Ainsi, le comptage des occurrences d'un k-itemset candidat est démarré dès que les compteurs de tous ses sous-itemsets de longueur $k-1$ sont suffisamment élevés vis-à-vis du seuil de support. Si le comptage du nouveau candidat débute sur le x-ième individu, il s'arrêtera sur le (x-1)-ième individu lors du passage suivant sur les données. Au final, *DIC* effectue moins de passages sur les données que *Apriori*.

L'algorithme *Partition* [SON95] est fondé sur un partitionnement des données en plusieurs corpus qui peuvent tenir en mémoire centrale. Pour qu'un itemset soit fréquent dans les données complètes, il doit l'être au moins sur un corpus. *Partition* analyse donc chaque corpus à la manière d'*Apriori*, mis à part que les supports des itemsets ne sont pas déterminés par comptage des occurrences. En effet, les corpus sont écrits dans un format mieux adapté au traitement en mémoire centrale, qui intervertit individus et variables : au lieu de décrire chaque individu par les items, on décrit chaque item par la liste des individus qui le vérifient, appelée *id-liste* (chaque individu est désigné par un identifiant *id*). Chaque itemset possède également son id-liste, et le nombre d'occurrences d'un itemset est donné par le nombre d'entrées dans son id-liste. Ainsi, pour déterminer le support d'un itemset i dans un corpus, il suffit de calculer l'intersection entre les id-listes de deux de ses sous-itemsets a et b tels que $i = a \wedge b$. Après avoir analysé tous les corpus, *Partition* vérifie pour chaque itemset qui a été trouvé fréquent sur un corpus s'il est bien fréquent dans les données complètes. En définitive, l'algorithme effectue deux passages sur les données complètes : un pour le partitionnement et un pour la vérification.

L'algorithme *FP-growth* [HPY00] a été développé pour traiter des données denses. Il travaille non pas directement sur les données mais sur une représentation condensée des données, appelée *FP-tree*. La représentation est construite en parcourant l'ensemble des itemsets en profondeur d'abord (contrairement à *Apriori* qui recherche en largeur). Cette étape de construction nécessite deux passages sur les données. Ensuite, l'algorithme génère tous les itemsets fréquents en mémoire centrale à partir de la représentation, si celle-ci peut tenir en mémoire.

Des comparaisons des principaux algorithmes d'extraction de règles d'association sont réalisées dans [HGN00] et [GZ03] sur des jeux de données réels et

synthétiques variés. Elles montrent que si les performances relatives des algorithmes peuvent varier selon les données étudiées, il n'existe pas un algorithme qui soit globalement meilleur que les autres.

3.4 Algorithmes à contraintes

3.4.1 Contraintes

En parallèle aux algorithmes exhaustifs ont été développés des algorithmes qui extraient les règles d'association "sous contraintes", c'est-à-dire avec des contraintes supplémentaires à celles du support et de la confiance. Avec des contraintes, l'utilisateur peut spécifier aux algorithmes les caractéristiques des règles qu'il recherche (s'il y a plusieurs contraintes, elles sont associées par conjonction). Dans le cadre de l'étude du panier de la ménagère, les contraintes peuvent être par exemple :

1. les règles doivent conclure sur un produit dont le prix vaut au moins σ,
2. le prix total des produits en prémisse doit être inférieur au prix total des produits en conclusion,
3. les règles ne doivent comporter que des produits textiles.

Ces exemples illustrent deux points importants. D'une part, les contraintes étudiées dans la littérature portent souvent sur une grandeur numérique qui caractérise les items (le prix des produits dans les exemples 1 et 2). D'autre part, les contraintes sont souvent exprimées à l'aide d'une hiérarchie de concepts décrivant les items ou une partie des items (dans l'exemple 3, une taxonomie des produits qui distingue entre autres les produits textiles).

Les algorithmes à contraintes utilisent les contraintes pour réduire l'espace de recherche à l'étape d'extraction des itemsets fréquents (l'étape de génération des règles reste identique aux algorithmes exhaustifs). Pour cela, ils traduisent les conjonctions de contraintes exprimées sur les règles en conjonctions de contraintes exprimées sur les itemsets[3]. Dans la suite, nous nous intéressons à l'étape d'extraction des itemsets fréquents, et considérons que les contraintes sont directement exprimées sur les itemsets.

Un algorithme à contraintes est optimisé pour une certaine classe de contraintes. Les deux principales classes de contraintes qui ont été étudiées sont les contraintes *anti-monotones* et *monotones*.

- Les contraintes anti-monotones sont définies par la propriété suivante : si un itemset vérifie une contrainte anti-monotone, alors tous ses sous-itemsets la vérifient aussi. Comme nous l'avons vu précédemment, la contrainte du seuil de support minimal est une contrainte anti-monotone.
- Les contraintes monotones sont définies par la propriété suivante : si un itemset vérifie une contrainte monotone, alors tous ses sur-itemsets la véri-

[3]Comme souligné dans [JB02], certaines contraintes sur les règles ne peuvent être traduites sur les itemsets (c'est le cas par exemple de la contrainte de confiance minimale). De telles contraintes ne sont donc pas utilisées pour réduire la taille de l'espace de recherche lors de l'extraction des itemsets fréquents. Les algorithmes les prennent en compte uniquement après la génération des règles, pour filtrer l'ensemble des règles obtenues.

fient aussi. Une contrainte monotone est donc la négation d'une contrainte anti-monotone (et vice versa).

Des exemples de contraintes sont données dans le tableau 3.1.

Contraintes monotones	**Contraintes anti-monotones**
$i \in S$	$i \notin S$
$S \supseteq \mathcal{I}$	$S \subseteq \mathcal{I}$
$\min_{i \in S}(x_i) \leq v$	$\min_{i \in S}(x_i) \geq v$
$\max_{i \in S}(x_i) \geq v$	$\max_{i \in S}(x_i) \leq v$
$\sum_{i \in S}(x_i) \geq v$ $(\forall i \in S, x_i \geq 0)$	$\sum_{i \in S}(x_i) \leq v$ $(\forall i \in S, x_i \geq 0)$
longueur$(S) \geq v$	longueur$(S) \leq v$
S est un itemset corrélé[4]	$support(S) \geq \sigma_{sp}$

i est un item, $\mathcal{I}$ un ensemble d'items, v est une valeur numérique, chaque item i possède une caractéristique numérique x_i.

TAB. 3.1 – Quelques contraintes monotones et anti-monotones sur un itemset S

3.4.2 Algorithmes

Les algorithmes à contraintes sont pour la plupart des généralisations d'*Apriori*. Ils parcourent l'ensemble des itemsets en largeur d'abord (notons tout de même que *FP-growth* a été adapté en algorithme à contraintes [PH00] [LLN02]). En fonction des contraintes utilisées, l'élagage engendré sur l'espace de recherche n'est pas le même. Les deux paragraphes ci-dessous expliquent comment les contraintes anti-monotones et monotones sont exploitées dans les algorithmes.

- Puisque *Apriori* est conçu pour tirer profit d'une contrainte anti-monotone, les contraintes anti-monotones peuvent être exploitées efficacement dans une structure d'algorithme issue d'*Apriori*. Une contrainte anti-monotone peut être prises en compte :
  - soit après la génération des itemsets candidats (entre les lignes 7 et 8 dans l'algorithme 3.1 page 76), si le test de la contrainte sur les itemsets ne nécessite pas de lire les données ;
  - soit après le passage sur les données (à la place de la ligne 13 dans l'algorithme 3.1 page 76), si le test de la contrainte sur les itemsets nécessite d'avoir lu les données (comme la contrainte de support minimal, qui est testée ligne 13).

- Les contraintes monotones sont exploitées lors de la génération des

[4] Un itemset est dit "corrélé" au seuil α si $P(\chi_p^2 > \chi_0^2) \leq 1 - \alpha$, où χ_0^2 est calculé sur la table de contingence croisant les variables qui apparaissent dans l'itemset (non pas les items binaires mais les variables d'origine qui, elles, peuvent être multimodales). Bien qu'il s'agisse d'ensembles non ordonnés d'items et non de liaisons orientées entre items, Brin *et al.* nomment ces itemsets "correlation rules" [BMS97] ou "dependence rules" [SBM98].

candidats (procédure `générerItemsetsCandidats()` à la ligne 6 de l'algorithme 3.1), mais elles sont moins évidentes à utiliser. Pour que les performances soient bonnes, il faut en effet que la contrainte permette d'énumérer l'ensemble des itemsets qui la vérifient uniquement à partir de la liste des items, sans avoir à consulter les individus dans les données. Nous nommons *fonction génératrice* (*member generating function* dans [NLHP98]) une procédure d'énumération adéquate. Par exemple, les contraintes syntaxiques[5] (définies uniquement à l'aide des items) disposent toutes d'une fonction génératrice (comme une procédure pour énumérer la liste des itemsets qui possèdent un item particulier), indépendamment du fait qu'elles soient monotones ou pas. Le problème est que les contraintes monotones ne disposent pas toutes d'une fonction génératrice. De plus, comme indiqué dans [JB02], une contrainte monotone peut amoindrir l'efficacité de l'élagage réalisé par les contraintes anti-monotones (voir exemple ci-dessous). Dans cette situation, l'introduction de la contrainte monotone dans le processus d'extraction des itemsets fréquents contribue à augmenter la taille de l'espace de recherche au lieu de la diminuer. Une approche qui peut même s'avérer plus efficace consiste à prendre en compte la contrainte en post-traitement, c'est-à-dire après l'extraction des itemsets fréquents, pour filtrer les résultats obtenus (degré zéro de l'extraction sous contraintes).

Exemple. Considérons des données décrites par un ensemble d'items $I = \{i_1, i_2, i_3\}$. La contrainte anti-monotone $\mathcal{C}_1$ est une contrainte de support minimal, et la contrainte monotone $\mathcal{C}_2$ est une contrainte qui exige que les itemsets contiennent l'item i_1. A la fin de la première itération de l'algorithme (niveau $k = 2$), l'ensemble des itemsets vérifiant les deux contraintes est $F_2 = \{i_1 \wedge i_2\ ,\ i_1 \wedge i_3\}$. L'itemset $i_2 \wedge i_3$ n'est pas dans F_2 puisque, comme il ne vérifie pas la contrainte $\mathcal{C}_2$, il n'a pas été généré comme itemset candidat. A l'itération suivante (niveau $k = 3$), l'ensemble des itemsets candidats $C_3 = \{i_1 \wedge i_2 \wedge i_3\}$ est généré, puis il doit être élagué. Supposons que $i_2 \wedge i_3$ ne vérifie pas $\mathcal{C}_1$. Cette information est inconnue par l'algorithme puisque cet itemset n'a pas été généré comme candidat au niveau inférieur. Il est donc impossible de prévoir que $i_1 \wedge i_2 \wedge i_3$ ne vérifie pas non plus $\mathcal{C}_1$. Dans le doute, $i_1 \wedge i_2 \wedge i_3$ est conservé dans C_3 et sera compté dans les données. Le problème réside donc dans le fait que certains itemsets peuvent être écartés par une contrainte monotone $\mathcal{C}_2$ alors qu'ils auraient permis à une contrainte anti-monotone $\mathcal{C}_1$ d'élaguer des pans entiers de l'espace de recherche. □

Ces principes généraux pour l'exploitation des contraintes se retrouvent dans différents travaux. L'algorithme *CAP* proposé dans [NLHP98] est optimisé pour des conjonctions de contraintes anti-monotones et de contraintes succintes (une classe de contraintes qui ont l'avantage de posséder une fonction génératrice). Dans [JB02] est présenté un algorithme générique qui exploite des conjonctions de contraintes anti-monotones et monotones. Plusieurs algorithmes sont également proposés dans [SVA97] pour des conjonctions ou disjonctions de contraintes syntaxiques qui imposent (contrainte monotone) ou interdisent (contrainte anti-monotone) la présence d'un item dans les itemsets. Les algo-

[5] Ces contraintes sont aussi appelées *métarègles* [KHC97], ou *templates* [KMR+94].

rithmes développés dans [BMS97] et [GLW00] exploitent la contrainte monotone des "itemsets corrélés"[4] avec d'autres contraintes (au moins une anti-monotone), mais ils n'extraient que les itemsets minimaux. Au final, malgré ces différents travaux, trouver la meilleure façon d'exploiter des combinaisons de contraintes anti-monotones et monotones quelles que soient les données reste un problème ouvert.

3.5 Quelle approche choisir ?

Si l'utilisateur veut exploiter des contraintes pour cibler un profil de règles qui l'intéresse, deux solutions s'offrent à lui :

1. exécuter un algorithme à contraintes ;
2. exécuter un algorithme exhaustif, puis appliquer les contraintes lors des post-traitements pour filtrer les règles produites [HG02] [GdB99].

La solution 1 présente deux avantages. Tout d'abord, l'utilisateur peut fixer des contraintes qui élaguent grandement l'espace de recherche, ce qui permet d'éviter de consommer des ressources et du temps pour des règles qui n'intéressent pas l'utilisateur. Si l'explosion combinatoire du nombre d'itemsets fréquents rend impossible l'exécution des algorithmes exhaustifs, la seule façon de pouvoir traiter les données avec le même seuil de support est ainsi d'essayer une approche par contraintes. Ensuite, puisque l'élagage ne repose pas que sur la contrainte de support minimal, il est possible avec les algorithmes à contraintes d'utiliser des seuils de support plus faibles (ou de traiter des données plus denses). Par exemple, des règles très spécifiques peuvent être découvertes par un algorithme à contraintes alors qu'elles n'auraient jamais pu être extraites par un algorithme exhaustif de type *Apriori*, à cause de l'explosion combinatoire. Les règles très spécifiques traduisent souvent des tendances dans les données totalement inconnues pour l'utilisateur, ce qui en fait pour lui des connaissances précieuses [Fre98].

Mais la solution 2 a aussi ses avantages. Le premier est que, nous l'avons vu, les algorithmes à contraintes ne peuvent pas toujours exploiter les contraintes non anti-monotones de manière optimale, si bien que prendre en compte ces contraintes après l'étape d'extraction des itemsets fréquents peut s'avérer plus rapide (ce qui revient à la solution 2). Le second avantage tient à la nature interactive et itérative de l'ECD. Lors d'un processus d'ECD, l'utilisateur effectue de multiples extractions successives, chaque nouvelle requête dépendant des résultats des extractions précédentes. Si un algorithme exhaustif parvient à s'exécuter jusqu'à son terme, alors tous les itemsets fréquents sont extraits une fois pour toute et disponibles pour toute requête ultérieure. Quelle que soit la succession d'extractions que l'utilisateur demandera par la suite, il suffira pour y répondre (du moment que le seuil de support n'est pas abaissé) de générer les règles à partir des itemsets déjà extraits, ce qui est l'étape rapide de l'extraction de règles d'association. En conséquence, à condition que l'étape d'extraction des itemsets fréquents puisse s'exécuter complètement, que l'utilisateur s'accommode de la durée de cette étape, et qu'il accepte de ne pas diminuer le seuil de support minimal par la suite, la solution 2 peut aussi favoriser les temps de réponses [HG02] [GdB99].

3.6 Conclusion

Les algorithmes qui ont été développés pour l'extraction de règles d'association se répartissent en deux catégories : les algorithmes exhaustifs et les algorithmes à contraintes. Dans l'absolu, aucune des deux approches ne surpasse clairement l'autre, les performances en temps et en mémoire dépendant des paramètres d'extraction et des données étudiées. Dans la pratique, un algorithme exhaustif est une solution convenable si les données étudiées sont creuses ou si l'utilisateur peut s'accommoder d'un temps de traitement long. Sinon, il n'y a pas de marche à suivre générale. La majorité des logiciels d'ECD implémentent uniquement des algorithmes exhaustifs.

Nous verrons au chapitre 6 que dans l'outil d'exploration de règles réalisé dans le cadre de cette thèse, nous avons opté pour des algorithmes à contraintes mais qui exploitent uniquement des classes très spécifiques de contraintes, qui élaguent grandement l'espace de recherche.

Visualisation interactive des règles : Proposition d'une méthodologie

4

"*Data graphics can do much more than simply substitute for small statistical tables. At their best, graphics are instruments for reasoning about quantitative information. Often the most effective way to describe, explore, and summarize a set of numbers –even a very large set– is to look at pictures of those numbers. Furthermore, of all methods for analyzing and communicating statistical information, well-designed data graphics are usually the simplest and at the same time the most powerful.*"

Edward R. Tufte (*The Visual Display of Quantitative Information*, 1983).

Sommaire

Les algorithmes de fouille de données produisent des règles d'association en si grandes quantités que l'utilisateur ne peut généralement pas les exploiter directement. En analyse du panier de la ménagère par exemple, il n'est pas rare d'obtenir des millions de règles portant sur plusieurs milliers d'items. Le post-traitement des résultats se révèle donc particulièrement crucial avec les règles d'association : de son efficacité à aider l'utilisateur à explorer cette masse d'information dépend la réussite de tout le processus ECD.

Dans les travaux dédiés au post-traitement des règles d'association, c'est souvent par visualisation des règles que s'effectue le post-traitement. Plus généralement, la visualisation est un moyen efficace d'introduire de la subjectivité dans chaque étape du processus ECD [FGW01]. Les représentations visuelles peuvent être exploitées :

- soit en tant que méthode de fouille de données à part entière, ce qui est souvent appelé *visual data mining* [Kei02] ;
- soit en collaboration avec des algorithmes de fouille de données pour faciliter et accélérer l'analyse des données étudiées, des résultats intermédiaires, ou des connaissances produites [Agg02] [Shn02] [HHC03].

La visualisation des règles d'association rentre dans ce dernier cas de figure.

Dans ce chapitre, nous proposons une méthodologie pour la visualisation interactive des règles d'association, conçue pour faciliter la tâche de l'utilisateur confronté à de grands ensembles de règles. Elle est fondée sur :

- des principes de visualisation d'information pour la construction de représentations visuelles efficaces [Ber67] ;
- des principes cognitifs de traitement de l'information dans le contexte des modèles de décision [Mon83].

Après un état de l'art sur le post-traitement des règles d'association, nous présentons en section 4.2 des travaux de visualisation d'information concernant l'élaboration d'outils de visualisation, et plus particulièrement la conception des représentations. La section 4.3 est consacrée plus particulièrement à deux tendances récentes de la visualisation d'information qui peuvent s'avérer utiles pour l'exploration de grands ensembles de règles : les représentations 3D et la réalité virtuelle. Enfin, nous étudions en section 4.4 les contraintes cognitives de l'utilisateur lors du post-traitement des règles d'association, puis proposons notre méthodologie pour la visualisation interactive des règles d'association.

4.1 Post-traitement des règles d'association : état de l'art

A la sortie des algorithmes de fouille de données, les ensembles de règles d'association sont de simples listes textuelles. Chaque règle consiste en un itemset qui constitue la prémisse, un itemset qui constitue la conclusion, et les valeurs numériques du support et de la confiance (voir chapitre 3). Les trois principales approches pour le post-traitement des règles d'association sont les suivantes :

1. évaluer, ordonner, et filtrer les règles avec des indices autres que le support et la confiance ;

2. organiser une exploration interactive des règles pour l'utilisateur ;
3. représenter les règles sous forme graphique.

La première partie de cette thèse est entièrement consacrée à l'approche 1. Ci-dessous, nous nous intéressons aux approches 2 et 3.

4.1.1 Exploration interactive des règles

Explorateurs de règles

Différents "explorateurs de règles" ont été développés pour aider l'utilisateur dans le post-traitement des règles. A l'instar des explorateurs de fichiers, ce sont des interfaces interactives qui présentent l'information sous forme textuelle. L'idée a d'abord été proposée par Mannila *et al.* [KMR+94], puis implémentée dans le logiciel TASA pour l'analyse des pannes d'un réseau de télécommunication [KMT96]. A l'aide de l'explorateur de règles intégré à TASA, l'utilisateur peut isoler les règles qui l'intéressent en ajustant des seuils sur des indices de règle et en spécifiant certaines contraintes syntaxiques (contraintes indiquant les items qui doivent apparaître ou ne pas apparaître dans les règles). Dans [MLW00], l'explorateur exploite un résumé de l'ensemble des règles obtenu par la méthode de [LHM99]. En sélectionnant une règle dans le résumé, l'utilisateur peut accéder aux règles plus spécifiques correspondantes. L'exploration des règles passe donc par la visite du résumé.

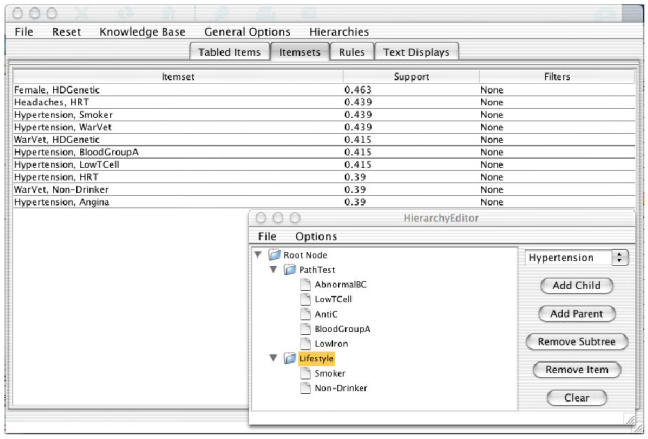


FIG. 4.1 – L'explorateur de règles IRSetNav [FR04]

Plus récemment, un explorateur de règles doté de nombreuses fonctionnalités a été présenté dans [FR04] (figure 4.1). Il permet de filtrer les règles par des contraintes syntaxiques plus ou moins générales, puisqu'elles peuvent prendre en compte une taxonomie des items. L'outil propose également à l'utilisateur de programmer les indices de règle de son choix pour trier et filtrer les règles. De plus, l'utilisateur a la possibilité de sauvegarder les règles qu'il juge intéressantes au fur et à mesure de son exploration. Un autre explorateur de règles est présenté

dans [TA02], mais il ne s'agit pas d'un outil générique. Il est destiné à l'analyse de données d'expression de gènes issues de puces à ADN, et repose sur un système de contraintes syntaxiques très complet pouvant prendre en compte une taxonomie des gènes.

Certains explorateurs de règles exploitent des indices de règle subjectifs, et peuvent ainsi tirer profit des connaissances de l'utilisateur sur les données. Par exemple, dans [LH96] et [LHWC99], l'explorateur ordonne les règles ou classe les règles en catégories selon qu'elles contredisent plus ou moins les croyances de l'utilisateur. Ces croyances sont exprimées différemment dans les deux approches.

- Dans [LH96], les croyances sont constituées des règles que l'utilisateur a jugées valides lors des explorations précédentes. Ces règles sont converties dans une représentation floue (*fuzzification*), sous le contrôle de l'utilisateur, pour plus de souplesse lors de la comparaison aux autres règles.
- Dans [LHWC99], une syntaxe est proposée pour que l'utilisateur puisse exprimer ses croyances avec différents degrés de précision. Une taxonomie des items peut être exploitée.

La principale limite des explorateurs de règles réside dans le mode de présentation textuel, qui ne convient pas à l'étude de grandes quantités de règles. Ces outils ont également le défaut de n'implémenter qu'un faible nombre d'indices de règle (trois maximum, souvent deux, mis à part [FR04]). Comme nous l'avons vu en partie 1, la qualité des règles se mesure pourtant selon de multiples points de vue.

Langages de requête

Le concept de *base de données inductive* a été introduit par Imielinski et Mannila dans leur article fondateur [IM96]. L'idée est d'enrichir les systèmes de gestion de bases de données pour qu'ils intègrent les méthodes de fouille de données. En d'autres termes, il s'agit de développer un langage de requêtes pour la fouille de données, généralisation de SQL qui permettrait de créer et manipuler les données mais également les connaissances extraites des données (des classifieurs, des ensembles de règles, les *clusters* issus de segmentations, etc.). Pour l'utilisateur, tout se passe comme s'il requêtait une base contenant à la fois données et connaissances, sans se soucier de savoir si les connaissances sont effectivement stockées dans la base ou bien générées dynamiquement à partir des données. Les bases de données inductives sont donc un projet ambitieux, mais difficile. Malgré les recherches effectuées depuis 1996 (bien souvent sur l'extraction de motifs fréquents –voir par exemple [Rae02] et [JB02]), de nombreux défis restent à relever. L'évaluation et l'optimisation des requêtes, en particulier, sont ardues. Avec les règles d'association par exemple, nous avons vu au chapitre 3 qu'il est difficile d'optimiser l'extraction des règles pour des contraintes qui, par essence, ne sont pas connues à l'avance (voir section 3.5 page 81).

Dans le cadre des bases de données inductives, plusieurs langages de requête ont été développés pour créer et manipuler des règles d'association, comme MSQL [IV99], DMQL [HFW+96], MINE RULE [MPC98], et XMINE [BCKL02]. Ces langages permettent d'effectuer l'extraction et/ou le post-traitement des

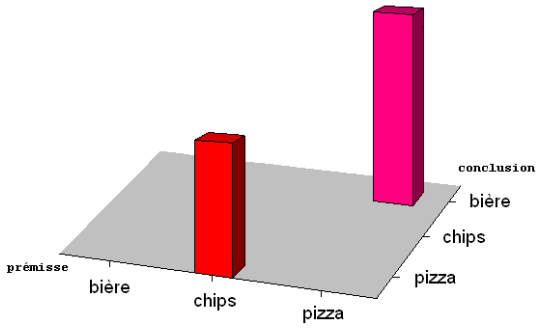


FIG. 4.2 – Une matrice itemset-à-itemset

règles d'association sous le contrôle de l'utilisateur. Toutefois, pour ce qui concerne le post-traitement des règles, les langages de requête sont peu pratiques (voir [BBMM02] pour une étude expérimentale), de la même façon que SQL utilisé seul ne convient pas à l'analyse de données.

4.1.2 Visualisation des règles

Les méthodes et outils de visualisation de règles décrits ci-dessous comportent généralement des fonctionnalités basiques de tri et de filtrage des règles, en fonction des items qui les constituent ou bien selon quelques indices de règle (peu d'indices en fait : le support, la confiance, et parfois une troisième mesure comme le lift).

Une première méthode de visualisation des règles d'association est la représentation par matrice. [HW01] et le groupe de recherche Quest[1] [AAB+96], ainsi que les logiciels DBMiner[2] [Han98], MineSet[3] [BQK97], Enterprise Miner[4], et DB2 Intelligent Miner Visualization[5], en donnent différentes implémentations. Dans une matrice itemset-à-itemset (figure 4.2), chaque colonne correspond à un itemset en prémisse et chaque ligne à un itemset en conclusion. Une règle entre deux itemsets est symbolisée dans la cellule à l'intersection par un objet 2D ou 3D dont les caractéristiques graphiques (généralement les dimensions et la couleur) représentent des indices de règle. Cette technique de visualisation a été améliorée en matrices item-à-règle [WWT99], où chaque ligne correspond à un item et chaque colonne à une règle (figure 4.3). La cellule à l'intersection d'un item et d'une règle est pleine ou vide suivant que l'item appartient ou non à la règle, la couleur de remplissage indiquant si l'item participe à la prémisse ou à la conclusion. La matrice est complétée par deux lignes qui indiquent le support et la confiance de chaque règle par la hauteur de barres dessinées en trois dimensions. Par rapport aux matrices itemset-à-itemset, les matrices item-à-règle sont moins encombrées et permettent une meilleure représentation des règles de plus de deux items. La principale limite de ces représentations matricielles est

[1] www.almaden.ibm.com/software/quest
[2] www.dbminer.com
[3] www.purpleinsight.com
[4] www.sas.com/technologies/analytics/datamining/miner/
[5] www.ibm.com/software/data/iminer/visualization/index.html

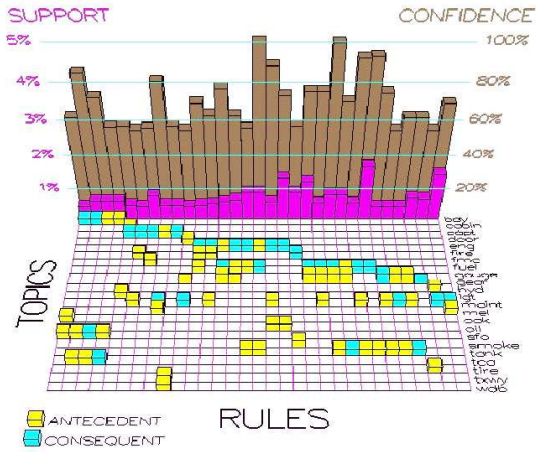


FIG. 4.3 – Une matrice item-à-règle dans [WWT99]

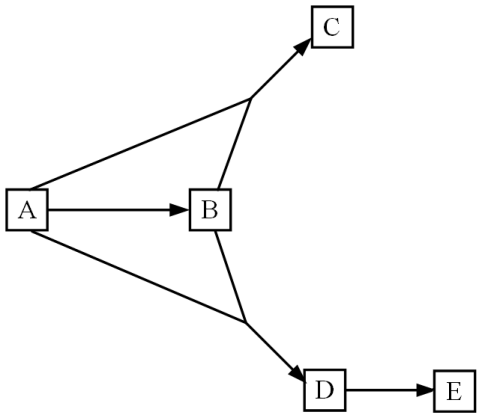


FIG. 4.4 – Un graphe d'items

qu'elles atteignent des tailles considérables dans le cas de grands ensembles de règles portant sur de nombreux items.

Les ensembles de règles d'association peuvent également être visualisés à l'aide d'un graphe[6] orienté (voir [KMR+94], [RR00], et les logiciels DBMiner[2] [Han98], CHIC[7] [CG05], et DB2 Intelligent Miner Visualization[5]). Dans ce type de représentations, les noeuds et les arcs symbolisent respectivement les items et les règles (voir figure 4.4 où les lettres désignent des items). Les indices de règle sont données par les arcs, par exemple avec la couleur ou l'épaisseur. Dans [HDH+], la méthode est implémentée en 3D avec un algorithme de type masses-

[6] Pour des règles de plus de deux items, il s'agit en fait d'un hypergraphe : les arcs peuvent contenir plusieurs branches pour relier plusieurs items en prémisse à plusieurs items en conclusion.

[7] www.ardm.asso.fr/CHIC.html

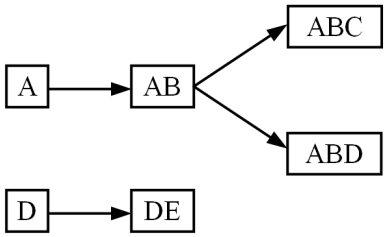


FIG. 4.5 – Un graphe d'itemsets

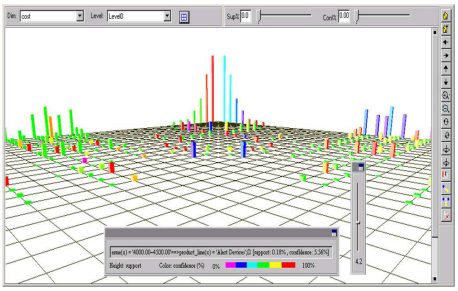
ressorts[8] qui optimise le placement des noeuds dans l'espace. Si la représentation par graphe a le mérite d'être très intuitive, elle admet deux principales limites. D'abord, elle fait implicitement apparaître les règles comme des relations transitives, alors que dans le cas général, les règles ne sont pas transitives (avec la plupart des indices de règle, la qualité des règles ne se propage pas par transitivité). Ensuite, elle ne convient pas non plus à la visualisation de grands ensembles de règles portant sur de nombreux items : le graphe est surchargé de noeuds et d'arcs qui se croisent, d'autant plus si des règles de plus de deux items sont considérées. En réponse à ce problème est proposée dans [Leh00] une représentation dynamique qui est un sous-graphe du treillis des itemsets. Dans ce graphe, les noeuds ne représentent pas les items mais les itemsets, de telle façon qu'une règle $(A \wedge B) \rightarrow (C)$ est symbolisée par un arc entre les noeuds $(A \wedge B)$ et $(A \wedge B \wedge C)$ (figure 4.5). Le graphe résultant est acyclique et comporte plus de noeuds mais moins de croisements d'arcs. L'utilisateur peut développer dynamiquement le graphe à sa guise en interagissant avec les noeuds.

Les autres méthodes de représentation de règles d'association ne concernent pas la visualisation de l'ensemble exhaustif des règles qui peuvent être extraites à partir d'un jeu de données. Etant donné quelques variables[9], ces méthodes ne représentent que le sous-ensemble des règles qui comportent uniquement ces variables. Elles permettent une étude approfondie d'un nombre limité de règles (la représentation devient rapidement inexploitable si trop de variables sont considérées), en particulier en montrant comment elles sont affectées par le changement des modalités des variables. Par exemple, la représentation dite en mosaïque pour les tables de contingence (*mosaic display*) a été adaptée aux règles d'association dans [HW01], chaque règle étant représentée par un rectangle dont l'aire est le support et la hauteur est la confiance (voir figure 4.7). Des techniques inspirées des coordonnées parallèles sont également utilisées pour visualiser des règles d'association [KT01] ou de classification[10] [HAC00]. Les

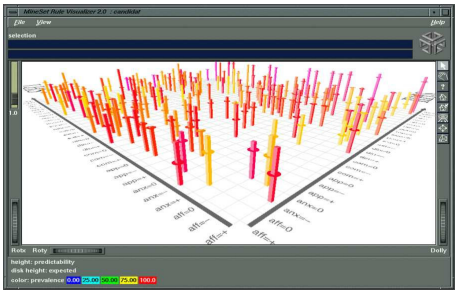
[8] Les noeuds sont considérés comme les masses, et la raideur du ressort entre deux masses est égal à la fréquence jointe (support) des deux items correspondant. Le graphe obtenu correspond à un état d'équilibre du système masses-ressorts. Ce type d'algorithme de dessin de graphe a été proposé initialement dans [Ead84].

[9] Nous rappelons que nous distinguons dans nos appellations les notions de *variables* et d'*items* (voir chapitre 1). Les variables sont les descripteurs qui se trouvent dans les données d'origine, elles peuvent être multimodales. Les items sont les descripteurs binaires issus du codage disjonctif des variables, chaque item correspondant à une modalité d'une variable.

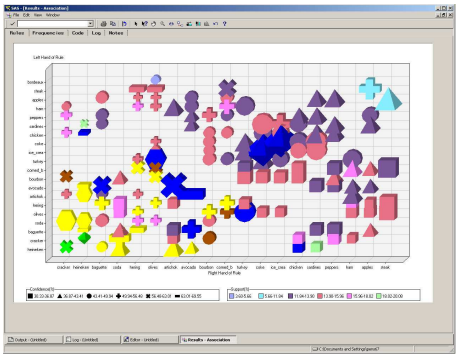
[10] Les règles de classification sont des règles qui concluent toutes sur la même variable. Contrairement aux règles d'association, les règles de classification sont extraites par des algorithmes supervisés.



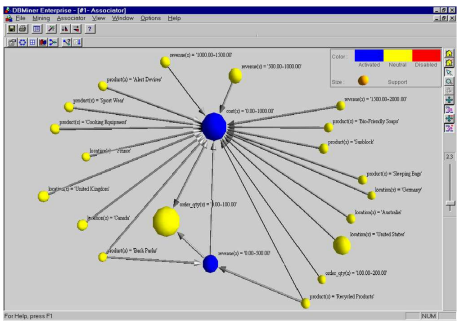
(a) DBMiner [Han98]



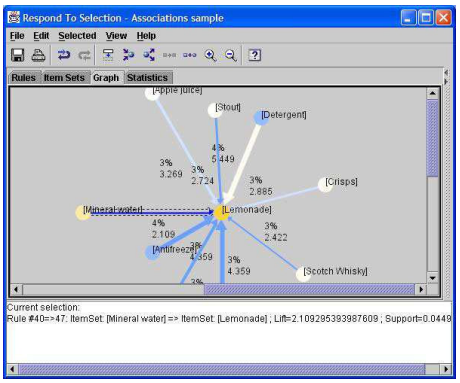
(b) Mineset[3]



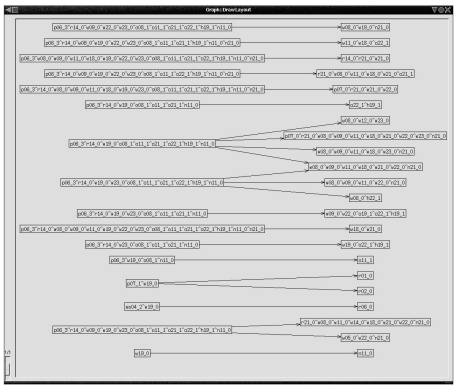
(c) Enterprise Miner[4]



(d) DBMiner [Han98]



(e) DB2 Intelligent Miner Visualization[5]



(f) PerformanSe-FELIX [Leh00]

FIG. 4.6 – Visualisation de règles d'association par matrices (a, b, c) ou graphes (d, e, f) dans quelques logiciels

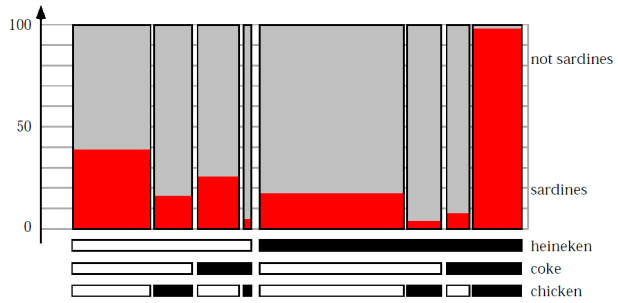


FIG. 4.7 – Représentation en mosaïque pour les règles (d'après [HW01])

Les règles représentées portent sur les items *heineken*, *coke*, et *chicken* en prémisse, et *sardines* en conclusion. La prémisse et la conclusion des règles se lisent respectivement sur l'axe horizontal et sur l'axe vertical. Le premier rectangle à gauche indique que la règle $(heineken = 0 \wedge coke = 0 \wedge chicken = 0) \rightarrow (sardines = 1)$ possède une confiance d'environ 40% (partie inférieure du rectangle), tandis que la règle contraire $(heineken = 0 \wedge coke = 0 \wedge chicken = 0) \rightarrow (sardines = 0)$ possède une confiance d'environ 60% (partie supérieure du rectangle). Le dernier rectangle à droite indique quant à lui une règle beaucoup plus forte : $(heineken = 1 \wedge coke = 1 \wedge chicken = 1) \rightarrow (sardines = 1)$ avec une confiance d'environ 100%. La règle contraire possède une confiance quasi-nulle.

variables sont représentées par des axes parallèles sur lesquels sont répartis les items, et chaque règle est symbolisée par une ligne brisée qui coupe les axes parallèles au niveau des items qu'elle contient. Les indices de règle peuvent être indiqués par l'épaisseur ou la couleur de la ligne, ou bien par des axes supplémentaires.

4.2 Visualisation d'information

4.2.1 Modélisation des outils de visualisation

Dans [CMS99], Card, Mackinlay, et Shneiderman proposent un modèle générique des outils informatiques de visualisation d'information. Il consiste en une suite de traitements interactifs permettant de passer des données en entrée à une visualisation en sortie : les transformations sur les entrées, puis l'encodage graphique, puis les transformations sur la vue (voir figure 4.8). Les données en entrée sont un ensemble d'individus décrits par des variables qui peuvent être *qualitatives nominales*, *qualitatives ordinales*, ou *quantitatives*. L'utilisateur contrôle chacune des transformations en interagissant avec l'outil de visualisation (directement dans l'interface où s'affiche la vue, ou bien dans une interface séparée).

1. Les transformations sur les entrées sont les traitements à effectuer sur les données avant qu'elles ne soient visualisées (sélection d'individus ou de

[11] www.inxight.com

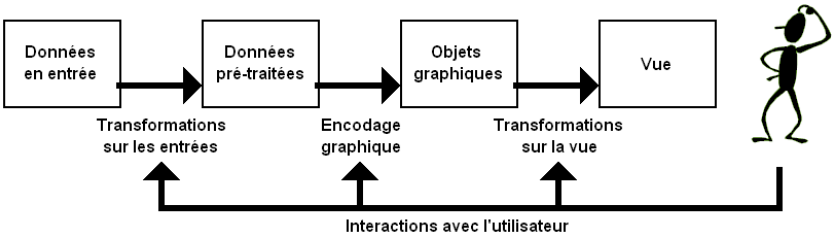
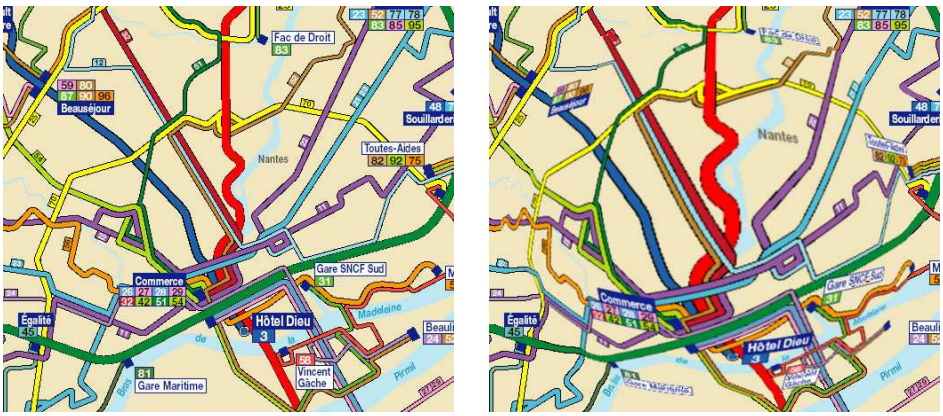


FIG. 4.8 – Modèle de [CMS99] pour la visualisation d'information



sans distorsion avec distorsion

FIG. 4.9 – Exemple de distorsion *fish eye*

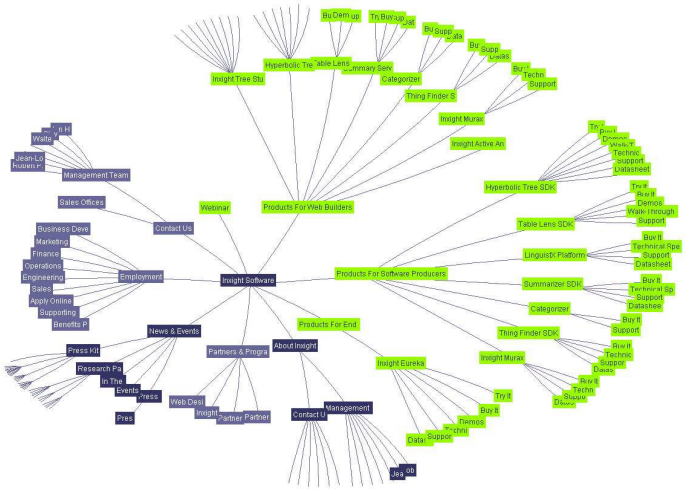


FIG. 4.10 – Visualisation d'un arbre dans un plan hyperbolique avec le logiciel Inxight StarTree[11]

variables, regroupement d'individus, formatage de variables, ordonnancement des modalités d'une variable, etc.). Pour ce qui concerne ces transformations, les techniques d'interaction les plus courantes sont les suivantes :
- le filtrage dynamique (introduit dans [WS92]), qui consiste à sélectionner les individus à visualiser par des requêtes dynamiques (l'affichage est mis à jour instantanément) portant sur les variables et soumises généralement par l'intermédiaire de cases à cocher (pour des variables qualitatives) ou de sliders (pour des variables quantitatives) [AS94b] ;
- le détail à la demande, qui consiste à choisir un élément dans la représentation et à faire apparaître des informations supplémentaires le concernant (souvent dans une fenêtre de type *pop-up* ou infobulle) ;
- le *brushing*, qui consiste à inclure dans ou exclure de la visualisation tout un sous-ensemble d'individus sélectionnés par l'utilisateur à l'aide du pointeur [Wil96].

2. L'encodage graphique est le coeur de la visualisation d'information : il s'agit de réécrire les données sous forme d'*objets graphiques* en associant à chaque variable dans les données une variable graphique (une position, une longueur, une aire, une couleur, une luminosité, une saturation[12], une forme, une texture, un angle, une courbure...). Les objets graphiques peuvent être de zéro à trois dimensions, c'est-à-dire un point, une ligne, une surface, ou un volume. L'évolution des objets dans le temps (modification des variables graphiques) peut constituer une dimension supplémentaire.
   Pour ce qui concerne l'encodage graphique, les interactions consistent à modifier les associations entre données et graphisme (comme par exemple changer de variable sur un axe). Une interface classique pour cela consiste à présenter à l'utilisateur un graphe dont les noeuds symbolisent les variables des données et les variables graphiques, et dont les arcs sont fixés par l'utilisateur pour indiquer les associations choisies.

3. Les transformations sur la vue concernent la présentation des objets graphiques à l'utilisateur. La vue affichée à l'écran peut être en 3D (même si les objets graphiques ne sont pas en 3D) ou en 2D. Pour ce qui concerne les transformations sur la vue, les techniques d'interaction les plus courantes sont les suivantes :
   - le contrôle du point de vue, qui s'effectue par translation, rotation, ou zoom, et peut être exocentrique (l'utilisateur déplace la représentation alors que le point de vue est fixe) ou bien égocentrique (l'utilisateur déplace le point de vue alors que la représentation est fixe) ;
   - les vues multiples (*overview+details*), qui permettent à l'utilisateur d'avoir une vue globale et une vue détaillée sur la même représentation par deux fenêtres (le brushing effectué dans une vue est aussi visible dans l'autre, c'est ce qu'on appelle le *linking+brushing*) [Shn96] ;
   - les techniques dites *focus+context* qui intègrent les détails dans la vue globale en les révélant autour du focus (point d'intérêt) de l'utilisateur,

[12] Ce que l'on appelle communément "couleur" regroupe en fait trois notions différentes : la couleur (est-ce rouge ? jaune ? bleu...), la luminosité (est-ce plus ou moins clair ?), et la saturation (est-ce plus ou moins intense ?) [Wil05].

soit par affichage direct (*fish eye* filtrant, introduit dans [Fur86]), soit par distorsion de la vue (*fish eye* déformant [SB92] et plans hyperboliques[13] [LRP95], voir figures 4.9 et 4.10).

Certaines techniques combinent plusieurs types de transformations. Par exemple, le zoom sémantique est un zoom (transformation sur la vue) qui change les données visualisées (transformation sur les entrées) : plus l'utilisateur zoome et plus le niveau de détail des données s'élève [HBL01]. Il est donc possible de zoomer continuellement tant que le niveau de détail le plus faible n'est pas atteint. Ce type d'outils de visualisation est appelé interfaces zoomables.

Au regard du modèle de Card, Mackinlay, et Shneiderman, nous pouvons formaliser les différences entre les méthodes de post-traitement de règles d'association ainsi :

- les méthodes d'exploration de règles (section 4.1.1) sont réduites aux transformations sur les entrées ;
- les méthodes de visualisation de règles (section 4.1.2) comportent bien sûr un encodage graphique et des transformations sur la vue, mais restent pauvres en ce qui concerne les transformations sur les entrées.

Comme nous le verrons à la section 4.5, la méthodologie proposée dans cette thèse pour le post-traitement de règles instancie le modèle dans ses trois composantes.

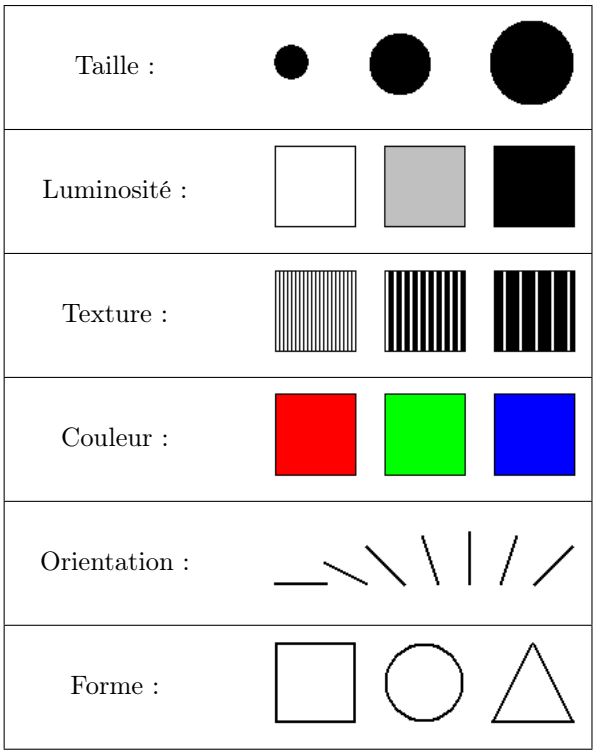
4.2.2 Principes de visualisation de Bertin

Plusieurs auteurs ont proposé des classifications des encodages graphiques dans le but de déterminer les encodages les plus efficaces en fonction des variables à représenter. Parmi ces travaux, ceux de Cleveland [CM84] puis Wilkinson [Wil05] font référence pour ce qui concerne les graphiques statistiques (diagrammes, nuages de points, etc.). Un second courant est issu de la cartographie, avec les travaux de MacEachren [Mac95] et surtout Bertin, dont la *Sémiologie graphique* [Ber67] est considérée comme la première théorie structurale des graphiques. Bien que Bertin couchait ses représentations sur du papier, ses principes font toujours référence pour la visualisation sur ordinateur. Pour ce qui concerne la visualisation, nous nous appuyons principalement dans cette thèse sur les travaux de Bertin.

Les variables graphiques étudiées par Bertin dans [Ber67] sont la position, la taille, la luminosité, la texture, la couleur, l'orientation, et la forme. La position

[13] Le plan hyperbolique est un plan possédant toutes les propriétés du plan euclidien mis à part que par un point extérieur à une droite on peut tracer plusieurs parallèles. Un tel plan peut être défini de la manière suivante : les points du plan hyperbolique sont les points intérieurs d'un disque (euclidien), et les droites du plan hyperbolique sont les cercles perpendiculaires au disque. L'intérêt des plans hyperboliques pour la visualisation est qu'ils permettent de représenter des entités non bornée (comme une droite) dans une surface de visualisation bornée.

[14] La variation d'orientation des traits est sélective s'il s'agit de traits isolés, ou bien de traits juxtaposés pour représenter des lignes (lignes hachurées), mais pas s'il s'agit de traits juxtaposés pour représenter des surfaces (surfaces hachurées).



TAB. 4.1 – Les variables rétiniennes de Bertin

Type de variable à représenter	variable nominale		variable ordinale	variable quantitative
Perception mise en jeu	associative	sélective	ordonnée	quantitative
Position	■	■	■	■
Taille		■	■	■
Luminosité		■	■	
Texture	■	■	■	
Couleur	■	■		
Orientation	■	■[14]		
Forme	■			

Les ■ indiquent quelles variables graphiques sont adaptées à la représentation de quelles variables dans les données.

TAB. 4.2 – Adéquation entre variables graphiques et variables à représenter (d'après Bertin [Ber67])

joue un rôle particulier en visualisation puisqu'il s'agit de l'information visuelle perceptivement dominante dans une représentation [Ber67] [CMS99] [Wil05]. Les six autres variables (voir tableau 4.1) sont appelées "variables rétiniennes" par Bertin car il est possible de percevoir leurs variations sans mettre à contribution les muscles du système optique, contrairement à la position. Pour ce qui concerne la taille, il est à noter que cette variable graphique désigne davantage des surfaces que des longueurs. En effet, comme l'écrit Bertin, "c'est la variation de surface qui constitue le stimuli sensible de la variation de taille" [Ber67]. Les encodages graphiques fondés sur des surfaces sont donc plus pertinents que les encodages fondés sur des longueurs. Remarquez que dans certains cas, la variation de surface se ramène à la variation d'une seule longueur (comme dans un diagramme en barres, où tous les rectangles ont un côté de même longueur).

Afin d'estimer les différentes possibilités d'encodage avec ces sept variables graphiques, Bertin identifie quatre attitudes envisageables pour une personne face à des données [Ber67] :

- la perception associative, lorsque l'utilisateur cherche à regrouper les différentes modalités d'une variable nominale pour pouvoir les repérer toutes ensemble ;
- la perception sélective, lorsque l'utilisateur cherche à distinguer les différentes modalités d'une variable nominale ;
- la perception ordonnée, lorsque l'utilisateur cherche à percevoir l'ordre des modalités d'une variable ordinale ;
- la perception quantitative, lorsque l'utilisateur cherche à percevoir les rapports entre les valeurs d'une variable quantitative.

Bertin synthétise ses principes de visualisation dans le tableau 4.2, qui montre l'adéquation entre variables graphiques et variables à représenter.

4.3 Visualisation d'information en 3D et en réalité virtuelle

4.3.1 2D ou 3D ?

Le choix entre représentations 2D et 3D en visualisation d'information reste une question ouverte [CMS99] [Che04]. Ceci s'explique en particulier par le fait que l'efficacité d'une visualisation dépende grandement de la tâche de l'utilisateur [CFB91]. De plus, si les représentations 3D sont bien souvent plus attrayantes, l'utilisation de la 2D a en sa faveur une longue et fructueuse expérience en visualisation d'information. D'une manière générale, les avantages de la 2D sur la 3D sont les suivants :

- Les représentations 2D sont plus faciles à appréhender que les représentations 3D. Il est en effet très facile de se "perdre" dans une représentation 3D (d'où l'intérêt de placer des points de repère dans la scène et de brider les possibilités de navigation de l'utilisateur) [Che04]. De plus, la perception de la profondeur dans une représentation 3D n'est pas triviale (par exemple, un objet de taille moyenne à l'écran représente-t-il un petit objet à l'avant de la scène ou un gros objet à l'arrière de la scène ?). Généralement, du fait de la profondeur, une représentation 3D comporte également

des occultations entre les objets graphiques qui la composent.
- Les capacités de traitement requises pour produire et faire évoluer une représentation sont généralement moins élevées dans le cas de la 2D que dans le cas de la 3D [HHC03]. En particulier, la simulation de la notion de profondeur dans une représentation 3D impose de calculer les perspectives, les incidences des ombres et des lumières, le rendu des textures, les occultations, etc.
- Les représentations 2D peuvent être explorées avec des interfaces classiques clavier/souris, alors que les interfaces de navigation dédiées aux scènes 3D (qui peuvent prendre en compte jusqu'à six degrés de liberté) sont moins répandues, beaucoup moins standardisées, et moins simples d'utilisation (voir [FMP01] pour un tour d'horizon). Les interfaces logicielles qui permettent de naviguer dans des scènes 3D à l'aide d'une simple souris sont généralement jugées peu efficaces.

Cependant, les avantages de la 3D sont les suivants :

- Alors qu'une représentation 2D est restreinte aux dimensions de l'écran, la dimension supplémentaire en 3D offre un point de vue vers l'infini, créant ainsi un large espace de travail pouvant contenir une grande quantité d'information[15] [CMS99]. Dans cet espace, les informations les plus importantes peuvent être placées à l'avant de la scène (objets les plus visibles) et ainsi mises en valeur par rapport aux informations moins importantes placées derrière (objets moins visibles). C'est pour cette raison que les représentations 3D sont parfois considérées comme des approches *focus+context*.
- Les objets d'une représentation 3D peuvent être plus complexes qu'en 2D. Ils possèdent plus de caractéristiques graphiques, et donc peuvent symboliser plus d'informations. En particulier, dans une représentation 2D, un objet n'est visible que sous un seul point de vue, alors qu'en 3D il est possible d'observer différentes faces du même objet.
- La navigation (contrôle du point de vue) dans une scène 3D est très intuitive, puisqu'il est possible de laisser l'utilisateur se "promener" dans la scène. L'utilisation de systèmes de visualisation immersifs comme un visiocube (*cave*) ou un visiocasque accentue encore davantage le caractère pseudo-naturel de la navigation de l'utilisateur [FMP01] [WS02].

Peu de recherches sont consacrées à la comparaison entre 2D et 3D. Pour ce qui concerne la visualisation statique (non interactive) de graphes statistiques, les représentations 3D (bien souvent des diagrammes en cylindres ou en parallélépipèdes à la place de diagrammes en barres) sont généralement déconseillées depuis les influentes publications de Tufte [Tuf83] et Cleveland [CM84]. Pourtant, les travaux de psychologie expérimentale de Spence [Spe90] et Carswell *et al.* [CFB91] montrent qu'il n'existe pas de différence significative de précision entre la 2D et la 3D pour la comparaison de grandeurs numériques. Sous certaines conditions, l'information est même traitée plus rapidement lorsqu'elle est représentée en 3D plutôt qu'en 2D [Spe90]. Pour la perception de tendances

[15] Toutefois, qu'il s'agisse de 2D ou de 3D, le nombre de pixels à l'écran est le même. Ainsi, une représentation 3D ne peut contenir plus d'objets qu'une représentation 2D qu'au prix d'occultations entre objets et d'une moindre résolution des objets éloignés.

générales dans les données (croissance ou décroissance), les résultats expérimentaux de [CFB91] font également état d'une amélioration des temps de réponse avec la 3D, mais au détriment de la précision.

D'autres travaux comparent 2D et 3D dans le cadre de la visualisation interactive. Dans [CM01], Cockburn et McKenzie s'intéressent au rangement et à la recherche de favoris (raccourcis vers des pages internet) dans un espace de visualisation 2D ou 3D. Avec l'interface 2D, les temps de traitement des utilisateurs sont meilleurs mais pas significativement. L'évaluation subjective des interfaces par les utilisateurs montre en revanche une préférence significative pour la 3D (ce que Spence [Spe90] et Carswell *et al.* [CFB91] pressentent également mais sans l'évaluer). Enfin, Ware et Franck [WF96] comparent visualisation statique de graphes 2D et visualisation interactive de graphes 3D, c'est-à-dire avec possibilité de changer le point de vue sur le graphe (l'idée à la base de ce choix est que l'interactivité est indispensable en 3D pour pourvoir élucider les occultations). Leurs travaux font état d'une augmentation significative de l'intelligibilité des graphes 3D par rapport aux graphes 2D. Plus précisément, leur expérience consiste à demander à des utilisateurs s'il existe un chemin de longueur deux entre deux noeuds choisis aléatoirement dans un graphe. Avec les graphes 3D, le taux d'erreur est diminué d'un facteur 2.2 pour des temps de réflexion comparables. Si l'on ajoute la vision stéréoscopique, le taux d'erreur est même diminué d'un facteur 3. On considère généralement que seule la stéréoscopie permet d'exploiter pleinement les caractéristiques des représentations 3D.

4.3.2 Réalité virtuelle

Un système de réalité virtuelle (RV) est un outil informatique de simulation qui permet à l'utilisateur de s'extraire virtuellement du monde réel pour changer de temps, de lieu, et d'interactions [FMP01] [CB03]. Il associe quatre composantes [FMP01] :

- l'interaction en temps réel
- avec un monde virtuel représenté en 3D
- par des interfaces sensorielles et/ou motrices
- pour permettre l'immersion pseudo-naturelle de l'utilisateur dans le monde virtuel.

L'immersion pseudo-naturelle de l'utilisateur signifie que l'utilisateur agit dans le monde virtuel comme il agirait dans le monde réel. C'est une notion en partie subjective. Un système de réalité virtuelle n'a cependant pas vocation à simuler le réel le plus fidèlement possible. Etant donnée la fonction que le système doit simuler (par exemple le pilotage d'un avion, un acte chirurgical, une opération de soudure), l'accent doit être mis sur les caractéristiques du système qui jouent un rôle important dans cette fonction [FMP01]. Ainsi, dans la simulation de la visite d'un musée, un ascenseur peut être modélisé grossièrement (par exemple avec une interaction de type menu, sans même "rentrer" dans l'ascenseur), alors que dans une simulation destinée à la thérapie des personnes claustrophobes, l'utilisation d'un ascenseur doit être simulée précisément.

En comparaison aux interfaces informatiques de base, les interfaces utilisées en réalité virtuelle sont moins répandues. Elles ont la particularité de pouvoir concerner tous les sens de l'utilisateur. Les interfaces sensorielles sont celles qui transmettent des stimulis sensoriels du système vers l'utilisateur : il peut s'agir par exemple d'enceintes, d'écrans, de cabines de projection, mais aussi de gants à retour tactile, de gants à retour thermique, de diffuseurs d'odeurs. Les interfaces motrices, quant à elles, transmettent des réponses motrices de l'utilisateur vers le système : capteurs de localisation, souris 3D, gants de données, bras à retour d'effort (*phantom*), tapis roulant. De nombreuses applications utilisent aussi leurs propres interfaces "artisanales".

Au delà du fait qu'il s'agisse d'une mauvaise traduction de l'Anglais "virtual reality[16]", l'expression "réalité virtuelle" est discutable car les environnements de réalité virtuelle ne simulent pas toujours la réalité mais peuvent aussi représenter des mondes imaginaires ou symboliques (le caractère pseudo-naturel de l'immersion relève alors de la métaphore). Par exemple, la réalité virtuelle commence à être utilisée en visualisation d'information, où elle permet d'explorer des données volumineuses et multi-dimensionnelles au moyen d'interfaces visuelles immersives. Ces interfaces procurent un grand champ visuel, généralement couplé à la vision stéréoscopique. Elles peuvent être [FMP01] [WS02] :

- un grand écran (stéréoscopie avec lunettes) ;
- une visiosalle, c'est-à-dire un système de projection sur écran semi-cylindrique ou sur trois écrans de la taille d'un mur (stéréoscopie avec lunettes) ;
- un visiocube (ou *cave*), c'est-à-dire une cabine cubique avec images projetées sur quatre ou six faces (stéréoscopie avec lunettes) ;
- un visiocasque, c'est-à-dire un casque équipé d'un capteur de localisation et de deux petits écrans à haute résolution pour la vision stéréoscopique.

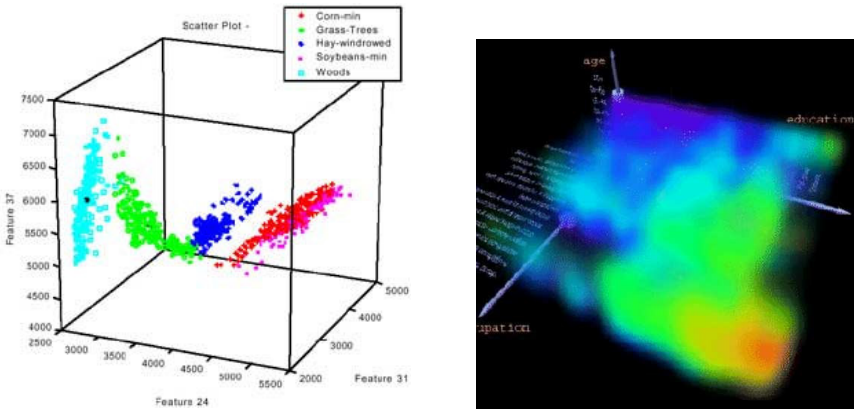
4.3.3 Applications 3D et RV en visualisation d'information

Une des représentations 3D les plus courantes en visualisation d'information est le nuage de points 3D (figure 4.11). Il est proposé dans de nombreux logiciels d'analyse de données. La principale innovation par rapport à la 2D est l'utilisation du rendu volumique, une technique classique en visualisation scientifique (en particulier en imagerie médicale) qui consiste à n'afficher que les voxels[17] qui représentent une certaine densité de matière. Cette technique a été adaptée aux nuages de points 3D dans [Bec97], faisant de l'opacité de chaque voxel une fonction de la densité de points (figure 4.11).

La 3D trouve aussi des applications en visualisation de graphes. Elle permet de représenter plus de noeuds, mais bien sûr uniquement au prix d'occultations. Les arbres coniques sont l'un des exemples les plus connus de cette approche. Ils ont été introduits dans [RMC91] pour la visualisation de structures hiérarchiques. Ce sont des arbres interactifs dessinés en 3D, verticalement (*cone trees*) ou horizontalement (*cam trees*, voir figure 4.12). Dans cette représentation, les

[16] Contrairement à sa version française, l'expression anglaise n'est pas un oxymoron. Une traduction plus adaptée aurait été par exemple "quasi-réalité" ou "pseudo-réalité" [Tis01].

[17] Le voxel est l'élément atomique d'une scène 3D, équivalent 3D du pixel.



sans rendu volumique avec rendu volumique

FIG. 4.11 – Exemples de nuages de points 3D

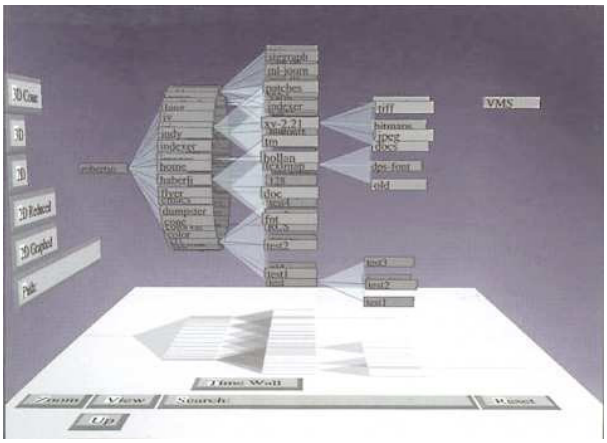


FIG. 4.12 – Arbre conique

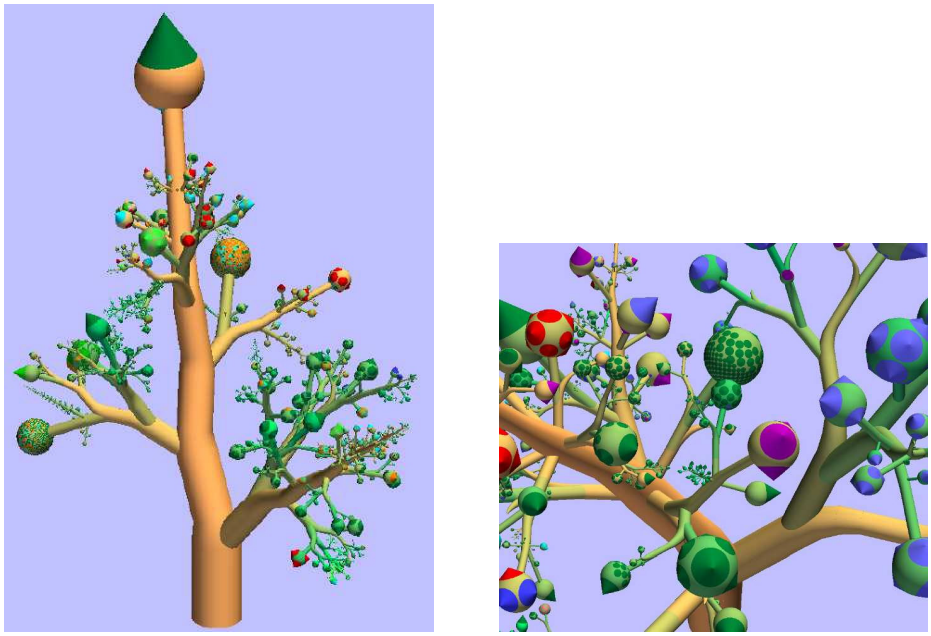


FIG. 4.13 – Visualisation d'un arbre par métaphore botanique

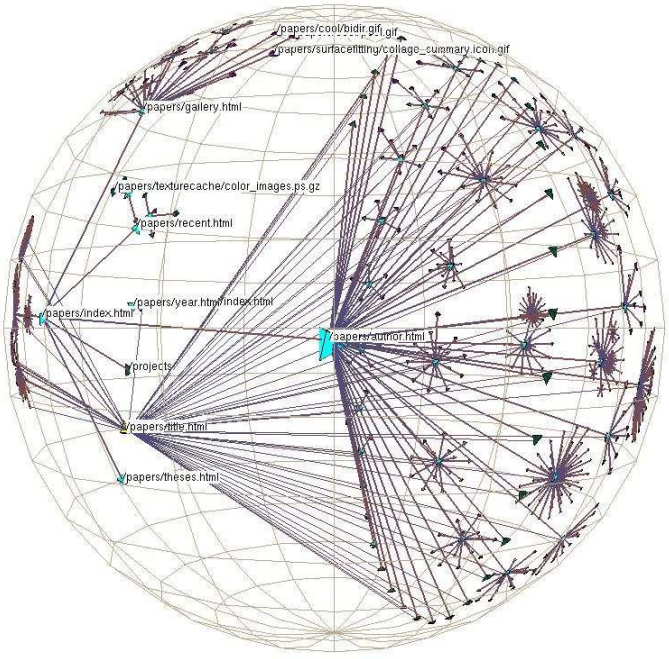


FIG. 4.14 – Visualisation d'un arbre dans un espace hyperbolique [Mun00]

noeuds fils sont placés circulairement autour du noeud père de façon à former un cône dont le diamètre décroît avec la profondeur de l'arbre. Quand l'utilisateur sélectionne un noeud dans l'arbre, une rotation amène le noeud au premier plan de la scène, ce qui permet d'explorer la hiérarchie. Une approche radicalement différente des arbres 3D est celle présentée dans [KvdWW01], qui repose sur une métaphore botanique (figure 4.13). Enfin, les travaux de Munzner sur les arbres et graphes 3D exploitent les propriétés des espaces hyperboliques, généralisation des plans hyperboliques[13] en 3D [Mun00]. Visuellement, les graphes dessinés dans ces espaces tiennent dans des sphères (voir l'exemple d'un arbre dans la figure 4.14). Il s'agit d'une approche *focus+context* : la partie du graphe située vers le centre de la sphère est grossie, tandis que les parties du graphe situées vers les bords de la sphère sont peu détaillées.

Les mondes virtuels (parfois appelés "cyber-espaces" [Che04]) constituent un autre courant important de la visualisation d'information en 3D. Il s'agit de représenter l'information par des objets répartis dans un grand espace affiché en 3D au sein duquel l'utilisateur peut naviguer (contrôler le point de vue). La navigation est d'ailleurs l'interaction fondamentale dans ce type de visualisation. La création de mondes virtuels en visualisation d'information repose généralement sur la métaphore de la *galaxie d'information* [Kro96] [CSBD97] ou bien sur la métaphore du *paysage d'information* [And95] [RCL+98] (figures 4.15 et 4.16). La différence entre les deux métaphores réside dans le fait que dans un paysage d'information, la dimension hauteur dans la position des objets ne peut pas être utilisée pour représenter une information. En effet, les objets sont posés sur un sol afin d'optimiser leur visibilité et de diminuer les occultations. Le sol peut être représenté pour faciliter la localisation des objets dans l'espace (en particulier, plus un objet est proche de la ligne d'horizon et plus il est éloigné). La galaxie d'information, quant à elle, ressemble plus à un nuage de points dans lequel les points (0D) sont remplacés par des objets (1D, 2D, ou 3D). Diverses applications ont montré l'efficacité des mondes virtuels pour l'exploration de grands corpus de données, dont les plus connues sont l'explorateur de fichiers FSN de Silicon Graphics (réutilisé dans MineSet[18] pour la visualisation des arbres de décision), l'explorateur de documents hypertexte Harmony [And95], et le système de visualisation d'Internet de Bray [Bra99] (voir figures 4.17 et 4.18).

Les premières applications de la réalité virtuelle en visualisation d'information font également leur apparition. Parmi elles, on trouve VRGobi, un système d'exploration de données multi-dimensionnelles en visiosalle [SCK+97]. Il implémente en version 3D les fonctionnalités de base du logiciel d'analyse de données XGobi[19]. VRGobi permet ainsi de projeter les données en un nuage de points 3D (avec *linking+brushing*), et de naviguer au sein de ce nuage. Le système TIDE [JL01] propose le même type de représentation (figure 4.19), mais il est fondé sur une architecture de travail collaboratif qui permet à plusieurs utilisateurs distants d'explorer ensemble les données. Placé dans un visiocube ou une visiosalle, chaque utilisateur peut "voir" les autres utilisateurs par le biais d'avatars, et discuter avec eux. Il peut également pointer une partie des données pour les montrer aux autres utilisateurs.

[18] www.purpleinsight.com
[19] www.research.att.com/areas/stat/xgobi/

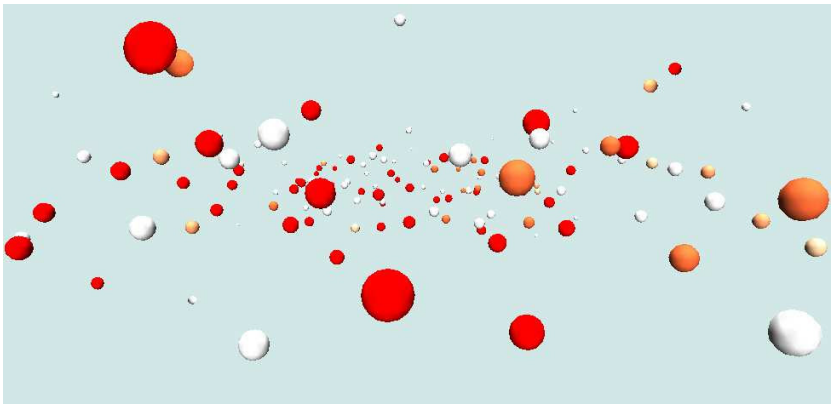


FIG. 4.15 – Exemple de visualisation avec la métaphore de la galaxie d'information

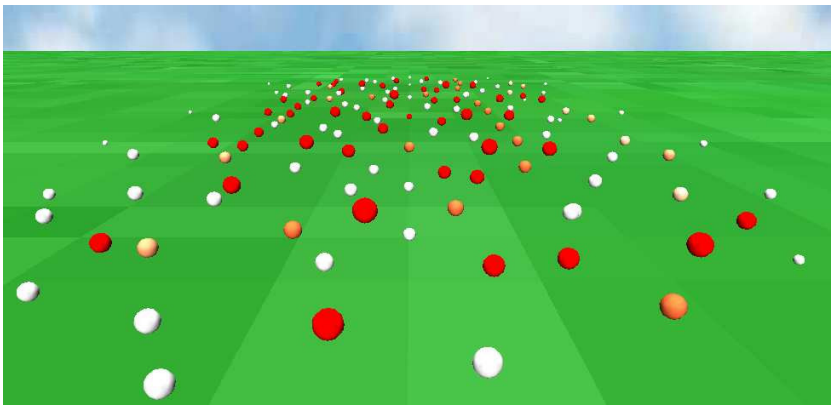


FIG. 4.16 – Exemple de visualisation avec la métaphore du paysage d'information

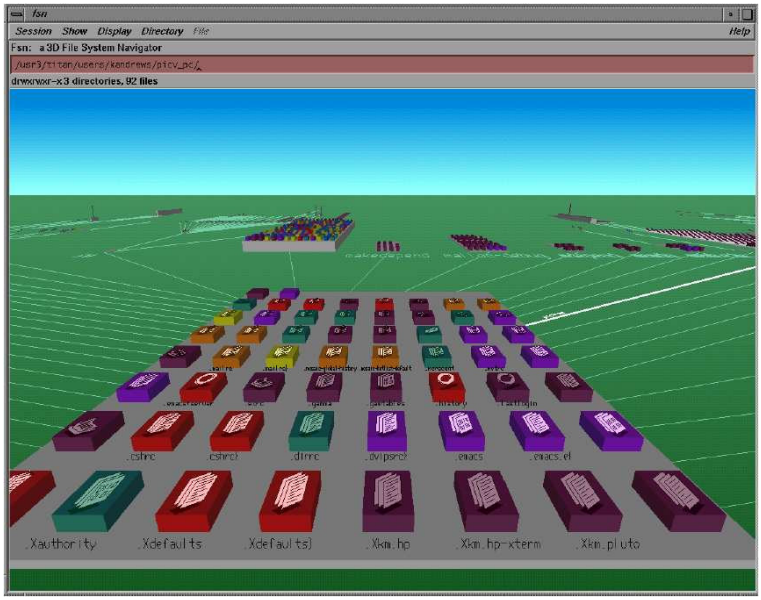


FIG. 4.17 – L'explorateur de fichiers FSN

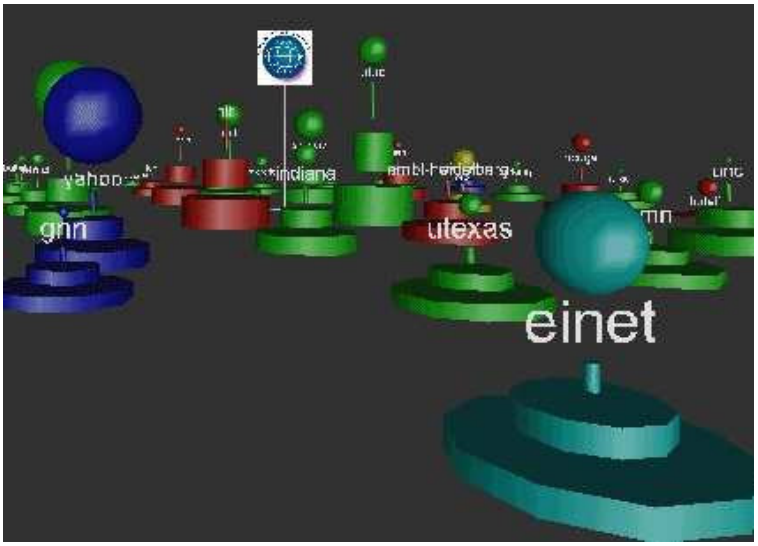


FIG. 4.18 – Visualisation de sites Internet dans [Bra99]

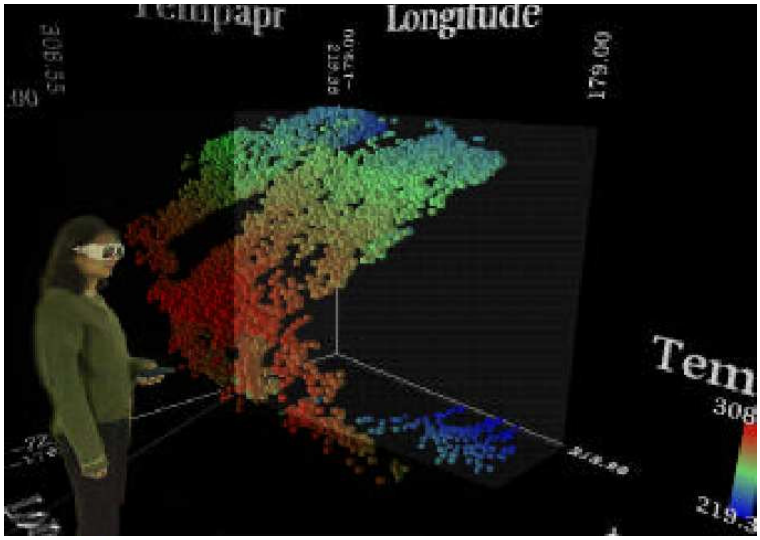


FIG. 4.19 – Exploration d'un nuage de points avec TIDE

FIG. 4.20 – Exploration d'un hypercube OLAP avec DIVE-ON

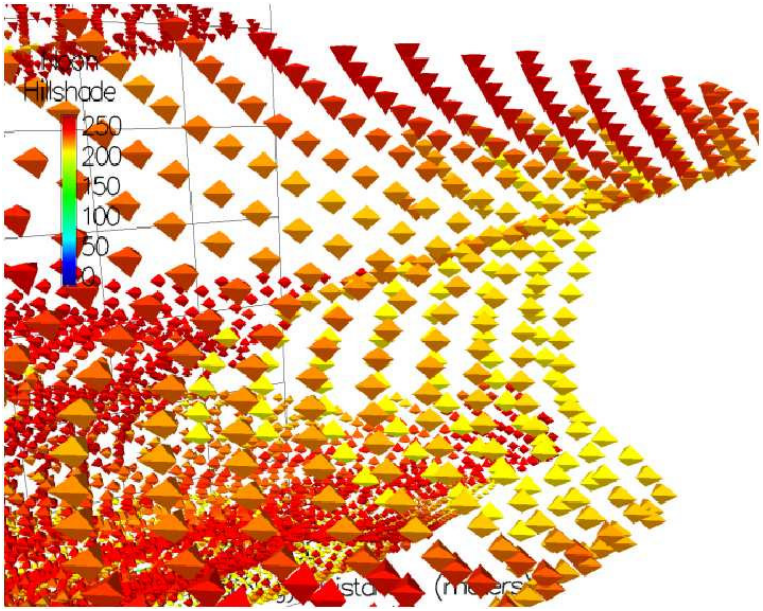


FIG. 4.21 – Exploration d'un nuage d'objets dans [NVGB03]

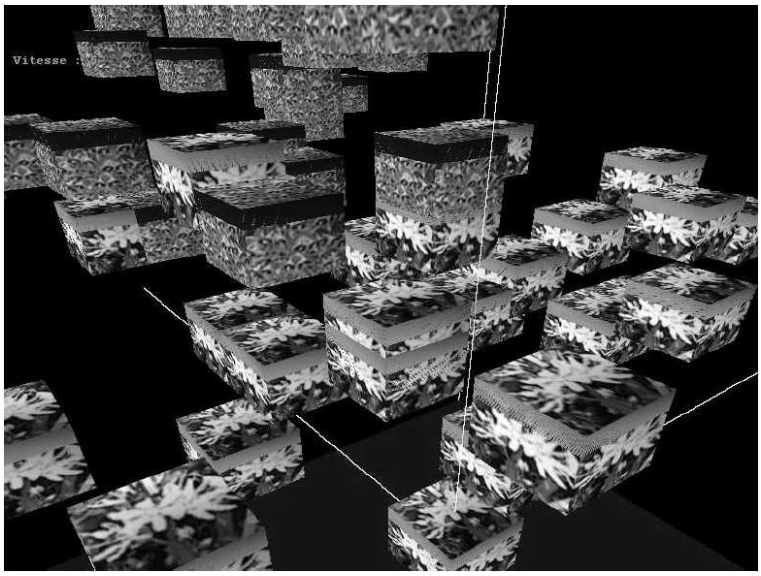


FIG. 4.22 – Exploration d'un nuage d'objets dans [AMG+03]

D'autres systèmes utilisent la métaphore de la galaxie. Par exemple, DIVE-ON [AZJ01] permet de visualiser un hypercube OLAP dans un visiocube, la position et la taille des objets représentant respectivement les dimensions et les faits de l'hypercube (figure 4.20). Trois dimensions seulement étant prises en compte pour placer les objets, l'hypercube est projeté en un cube. L'utilisateur peut naviguer au sein du cube et appliquer les opérateurs classiques de l'analyse OLAP. Deux autres approches exploitant la métaphore de la galaxie sont proposées dans [NVGB03][20] et [AMG+03]. Ce sont des systèmes d'exploration de données qui affichent en stéréoscopie un nuage d'objets au sein duquel l'utilisateur peut naviguer (figures 4.21 et 4.22). Conçu plus particulièrement pour l'exploration de données multimédia, le système de [AMG+03] supporte également l'affichage d'images et la lecture de sons et de vidéos. Si l'approche de [NVGB03] est testée dans des équipements coûteux comme un visiocube et une visiosalle, la solution présentée dans [AMG+03] a été développée dans un souci de minimisation du prix. En particulier, l'interface visuelle utilisée est un moniteur d'ordinateur classique, l'accent étant mis davantage sur les interfaces motrices (capteur de localisation et gant de données). Au final, cette solution montre qu'il est aussi possible de réaliser une application de réalité virtuelle pour un coût abordable (moins de 10 000 euros dixit les auteurs).

4.4 Contraintes cognitives de l'utilisateur lors du post-traitement des règles

4.4.1 Tâche de l'utilisateur

Lors du post-traitement des règles d'association, l'utilisateur se trouve face à de grands ensembles de règles décrites par des indices. La tâche de l'utilisateur consiste alors à fouiller les règles pour trouver des connaissances intéressantes pour la prise de décision. Pour cela, il a besoin d'interpréter les règles dans sa sémantique métier et d'évaluer leur qualité. Les deux indicateurs pour la prise de décision sont donc la syntaxe des règles (les items qui prennent part dans chaque règle) et les indices.

La tâche de l'utilisateur est difficile pour deux raisons. Tout d'abord, la profusion de règles à la sortie des algorithmes de fouille de données interdit toute exploration exhaustive. Ensuite, de par leur nature non supervisée, les règles d'association sont typiquement utilisées lorsque l'utilisateur ne connaît pas assez précisément ce qu'il recherche pour pouvoir l'exprimer dans la terminologie des données. Il ne peut donc pas spontanément formuler des contraintes qui isoleraient directement les règles qui l'intéressent.

4.4.2 Hypothèses sur le traitement cognitif de l'information

D'après l'hypothèse de rationalité limitée (*bounded rationality*) [Sim79], un processus de prise de décision peut être vu comme la recherche d'une structure de

[20] 3DVDM : www.idi.ntnu.no/~hrn/3dvdm/ et www.cvmt.dk/projects/3dvdm/

dominance. Plus précisément, l'utilisateur confronté à un ensemble d'options décrites par plusieurs attributs essaie de trouver une option qu'il considère comme dominante, c'est-à-dire une option qu'il juge meilleure que les autres d'après sa représentation courante du contexte de décision [Mon83]. Ce modèle de prise de décision peut être utilisé pour le post-traitement des règles d'association en considérant les règles comme un type particulier d'options, avec les items et les indices de règle comme attributs. Selon Montgomery [Mon83], l'utilisateur isole un nombre limité d'options potentiellement intéressantes et réalise des comparaisons entre elles. Ceci s'effectue à de multiples reprises durant le processus de prise de décision. En particulier, il souligne que "le processus de prise de décision acquiert une certaine direction dans le sens où plusieurs options et attributs reçoivent davantage d'attention que les autres [...]. La direction du processus peut être donnée plus ou moins consciemment. Des changements de direction peuvent avoir lieu plusieurs fois durant le processus, particulièrement quand l'utilisateur peine à trouver une structure de dominance."

Par ailleurs, dans [Ban94] est proposée une méthodologie d'ECD appelée *attribute focusing* qui s'appuie sur des résultats expérimentaux concernant le comportement de l'utilisateur pendant le processus d'ECD. Cette méthodologie est fondée sur un filtre automatique qui détecte par des mesures statistiques un petit nombre d'attributs potentiellement intéressants. Ce filtre guide l'attention de l'utilisateur sur un sous-ensemble des données de taille réduite et donc plus intelligible. L'intérêt de cibler un petit nombre d'attributs pour le traitement cognitif de l'information a aussi été grandement confirmé par les travaux sur les stratégies de décision (voir par exemple l'heuristique de la base mobile dans [BM92]). En effet, du fait de ses capacités cognitives limitées, l'utilisateur n'examine à chaque instant qu'une petite quantité d'information.

De ces différents travaux sur le traitement cognitif de l'information, nous établissons trois principes sur lesquels repose notre méthodologie pour le post-traitement des règles d'association :

P1. permettre à l'utilisateur de cibler son attention sur un sous-ensemble limité de règles, décrit par un petit nombre d'attributs (items et indices de règle) ;

P2. permettre à l'utilisateur d'effectuer des comparaisons entre les règles du sous-ensemble ;

P3. permettre à l'utilisateur de changer de sous-ensemble de règles à n'importe quel moment pendant le post-traitement.

4.5 La méthodologie *Rule Focusing* pour la visualisation interactive des règles

Notons $\mathcal{R}$ l'ensemble complet des règles produites par un algorithme exhaustif d'extraction de règles d'association. Notre méthodologie pour le post-traitement des règles d'association, nommée méthodologie *RF* (pour *Rule Focusing*), est conçue pour faciliter la tâche de l'utilisateur confronté à des ensembles $\mathcal{R}$ de grande taille. Elle consiste à laisser l'utilisateur naviguer dans $\mathcal{R}$ à sa guise en explorant des sous-ensembles limités de règles au moyen d'une représentation graphique des règles et de leurs indices. En d'autres termes, l'utilisateur dirige

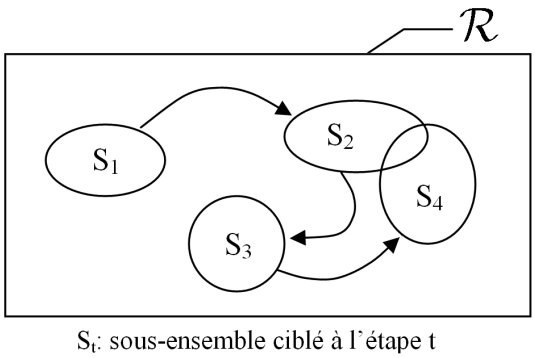


FIG. 4.23 – Des explorations locales successives dans l'ensemble $\mathcal{R}$ des règles

par tâtonnements successifs une **suite d'explorations locales visuelles** en fonction de son intérêt pour les règles (figure 4.23). Ainsi, l'ensemble $\mathcal{R}$ est exploré sous-ensemble après sous-ensemble, de telle façon que l'utilisateur n'ait jamais à l'appréhender dans sa globalité. A chaque étape de la navigation[21], l'utilisateur doit prendre une décision pour choisir le prochain sous-ensemble à visiter. C'est par là même que s'exprime la subjectivité de l'utilisateur dans le post-traitement des règles.

La méthodologie *RF* intègre les principes cognitifs de la section 4.4.2 de la manière suivante.

- Des relations permettent de cibler les sous-ensembles et naviguer entre eux (principes P1 et P3). Nous les nommons *relations de voisinage*.
- L'utilisateur visualise les sous-ensembles pour les visiter (principe P2). Entre autres facilitations cognitives, la visualisation simplifie en effet les comparaisons entre informations quand l'encodage graphique est bien choisi [CRC03].

Les relations de voisinage et la méthode de visualisation doivent prendre en compte les deux indicateurs impliqués dans la tâche de l'utilisateur : la syntaxe des règles et les valeurs des indices (voir section 4.4.1).

4.5.1 Relations de voisinage

Les relations de voisinage définissent comment les sous-ensembles de règles sont constitués (principe cognitif P1) et comment l'utilisateur peut passer d'un sous-ensemble à un autre (principe P3). En tant que vecteurs de la navigation pour l'utilisateur, ces relations sont un élément fondamental de la méthodologie *RF*. Elles sont définies de la manière suivante : dans l'ensemble des règles $\mathcal{R}$, une relation de voisinage associe chaque règle à un sous-ensemble limité de règles, qui sont qualifiées de voisines (figure 4.24). Donc avec x relations, l'utilisateur peut atteindre x sous-ensembles de règles voisines à partir d'une règle, et à partir d'un sous-ensemble contenant y règles, il peut atteindre $x.y$ sous-ensembles

[21] Nous appelons "navigation" le fait de passer d'un sous-ensemble à un autre, alors que "exploration" se réfère au processus entier supervisé par notre méthodologie, c'est-à-dire la navigation parmi les sous-ensembles et les visites des sous-ensembles (explorations locales).

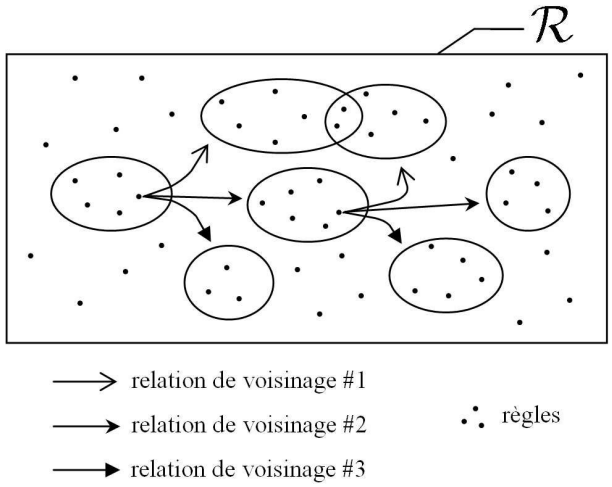


FIG. 4.24 – Une relation de voisinage associe chaque règle à un sous-ensemble de règles

voisins possibles. Pour naviguer d'un sous-ensemble à un autre, l'utilisateur doit effectuer deux choix : quelle relation de voisinage appliquer, et sur quelle règle.

D'un point de vue mathématique, les relations de voisinage sont des relations binaires dans l'ensemble $\mathcal{R}$ des règles (relations non symétriques dans le cas général). Toujours dans le but de faciliter la tâche de l'utilisateur, nous choisissons des relations de voisinage qui font sens pour l'utilisateur :

$$\forall (r_1, r_2) \in \mathcal{R}^2, (r_1) est\_voisin\_de(r_2) \Leftrightarrow \left(\begin{array}{l} \text{l'utilisateur juge que } r_1 \text{ est proche} \\ \text{de } r_2 \text{ selon un certain point de vue} \end{array} \right)$$

Toute relation qui possède une sémantique pertinente pour l'utilisateur peut être envisagée comme relation de voisinage. Conséquemment, avant de débuter le post-traitement des règles, la participation de l'utilisateur est nécessaire pour définir les relations à utiliser.

Voici par exemple quatre relations de voisinage $(r_1) est\_voisin\_de(r_2)$ envisageables :

1. r_1 est voisin de r_2 si et seulement si r_1 et r_2 possèdent la même conclusion ;
2. r_1 est voisin de r_2 si et seulement si r_1 est une exception[22] de r_2 ;
3. r_1 est voisin de r_2 si et seulement si r_1 possède une prémisse plus générale que celle de r_2 (la figure 4.25 illustre un graphe de cette relation de voisinage) ;
4. r_1 est voisin de r_2 si et seulement si r_1 possède le même support et la même confiance que r_2 à 0.05 près.

[22]Les exceptions d'une règle r sont toutes les règles qui possèdent la même prémisse que r mais augmentée d'un item supplémentaire, et qui concluent sur la même variable que r mais pas sur le même item. Par exemple, la règle $(jour = vendredi) \rightarrow (déjeuner = poisson)$ admet pour exception $(jour = vendredi \wedge date = 25décembre) \rightarrow (déjeuner = dinde)$.

Les relations de voisinage 1, 2, et 3 sont fondées sur la syntaxe des règles, tandis que la relation 4 repose sur deux indices de règle. Par ailleurs, la relation 1 est une relation d'équivalence, alors que la relation 2 n'est ni réflexive, ni symétrique, ni transitive. La relation 3 est uniquement transitive, et la relation 4 est réflexive et symétrique mais pas transitive.

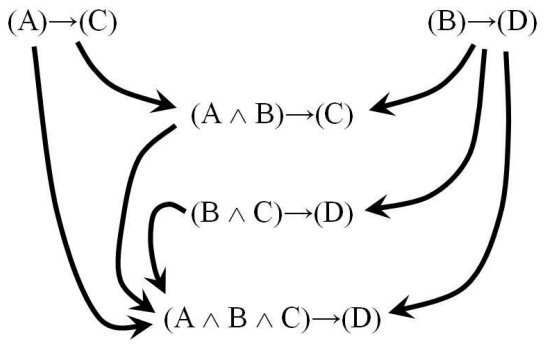


FIG. 4.25 – Graphe de la relation 3
pour un ensemble $\mathcal{R}$ de cinq règles (les lettres désignent des items)

L'originalité de la méthodologie *RF*, en comparaison avec les autres méthodes d'exploration de règles (décrites section 4.1.1), réside principalement dans le concept de relation de voisinage. Avec un explorateur de règles ou un langage de requêtes, l'utilisateur peut atteindre n'importe quel sous-ensemble de règles mais à condition qu'il puisse spécifier explicitement les contraintes qui le délimitent. Avec la méthodologie *RF*, la spécification des contraintes est implicite puisque intégrée dans les relations de voisinage, qui peuvent être vues comme des classes de contraintes. Nous pensons que l'application de relations de voisinage facilite davantage la tâche de l'utilisateur que la spécification explicite de contraintes.

Exemple. Considérons la scène suivante dans un processus de post-traitement de règles.

L'utilisateur trouve une règle intéressante $(A \wedge B \wedge C) \rightarrow (D)$*, où les lettres désignent des items. Il pense que la combinaison de ces quatre items est pertinente mais souhaiterait changer l'ordre des items pour vérifier si les règles* $(A \wedge B \wedge D) \rightarrow (C)$*,* $(A \wedge C \wedge D) \rightarrow (B)$*, et* $(B \wedge C \wedge D) \rightarrow (A)$ *ne sont pas mieux évaluées par les indices (et pourquoi pas également les règles* $(A \wedge B) \rightarrow (C \wedge D)$*,* $(A \wedge C) \rightarrow (B \wedge D)$*...).*

Avec la méthodologie *RF*, l'utilisateur peut accomplir ce scénario en une seule interaction. Pour cela, il a uniquement besoin de la relation de voisinage "r_1 est voisin de r_2 si et seulement si r_1 et r_2 possèdent les mêmes items". En revanche, si l'utilisateur emploie un langage de requêtes pour réaliser le post-traitement des règles, alors il doit taper une requête ad hoc pour récupérer chacune des règles désirées. S'il utilise un explorateur de règles, alors il doit retrouver chaque règle manuellement par l'interface graphique. Dans les deux cas, la tâche de l'utilisateur peut s'avérer fastidieuse. □

4.5.2 Visualisation des règles

Dans la méthodologie *RF*, l'utilisateur visualise les règles pour pouvoir plus facilement les comparer entre elles (principe cognitif P2). Nous reprenons le modèle de Card, Mackinlay, et Shneiderman (voir section 4.2.1 page 92) pour décrire cette visualisation. Les données en entrée du modèle sont ici l'ensemble de règles $\mathcal{R}$.

Transformations sur les entrées
Nous profitons du fait que l'utilisateur concentre son attention sur des sous-ensembles pour ne visualiser que le sous-ensemble en cours d'exploration. Ainsi, ce sont les relations de voisinage qui réalisent les transformations sur les entrées en ciblant les sous-ensembles. Des opérateurs interactifs doivent être intégrés à la visualisation pour que l'utilisateur puisse "activer" les relations.

Encodage graphique
Un défaut important des méthodes de visualisation exposées dans la section 4.1.2 est qu'elles mettent très peu en valeur les indices de règle, alors même que ce sont des informations primordiales pour la validation des règles. Par exemple, les méthodes de visualisation par matrice, par graphe, et par coordonnées parallèles utilisent la couleur pour représenter certains indices de règle, alors que ce choix d'encodage graphique pour des variables quantitatives est connu pour être mauvais en visualisation d'information[23] [Ber67] [Tuf83]. Dans la méthodologie *RF*, pour encoder les indices de règle de manière pertinente, nous nous référons aux principes de visualisation de Bertin [Ber67] (voir section 4.2.2 page 94). En ce qui concerne les variables quantitatives, ils indiquent qu'un encodage efficace doit s'effectuer avec des positions ou des tailles. En particulier, la position est l'information visuelle perceptivement dominante dans une représentation [Ber67] [CMS99] [Wil05] ; elle devrait être utilisée pour les indices les plus importants.
Une autre limite des méthodes de visualisation exposées dans la section 4.1.2 est le faible nombre d'indices de règle qu'elles encodent (trois maximum, souvent deux). Comme nous l'avons vu en partie 1, la qualité des règles se mesure pourtant selon de multiples points de vue. Pour la méthodologie *RF*, nous nous référons à la classification 1.16 de la page 39 et proposons d'encoder au moins quatre indices : un indice descriptif d'écart à l'équilibre, un indice descriptif d'écart à l'indépendance, un indice statistique d'écart à l'équilibre, et un indice statistique d'écart à l'indépendance. Comme nous l'avons indiqué au chapitre 1, un tel quadruplet d'indices permet selon nous de mesurer quatre aspects fortement "orthogonaux" de la qualité des règles.

[23] Faute d'encodage aussi attractif, l'utilisation de la couleur pour représenter des variables qualitatives ordinales est très souvent "tolérée" quand les modalités sont peu nombreuses [Spe00] [Wil05]. Ceci peut en particulier s'appliquer à des variables quantitatives comme les indices de règle si l'on accepte de les discrétiser. En toute rigueur, une telle solution est à rejeter puisque les couleurs n'induisent aucun ordre universel, leur ordonnancement est propre à chacun [Ber67] [Tuf83].

Transformations sur la vue
Les travaux en perception visuelle montrent que l'être humain a une perception d'abord globale d'une scène, avant de porter son attention sur les détails [HBL01] (c'est ce qui a motivé le développement des approches *overview+details* et *focus+context* décrites à la section 4.2.1). Ceci se retrouve en particulier dans une formule bien connue énoncée par Shneiderman et considérée le mantra de la visualisation d'information : "overview first, zoom and filter, then details on demand" [Shn96]. Dans la méthodologie *RF*, l'utilisateur doit donc pouvoir passer aisément d'une vue globale à une vue détaillée en interagissant avec la visualisation.

4.6 Conclusion

Dans ce chapitre, nous avons présenté la méthodologie *Rule Focusing* (*RF*) pour la visualisation interactive des règles d'association. Elle est conçue pour faciliter la tâche de l'utilisateur confronté à de grands ensembles de règles en prenant en compte ses capacités de traitement de l'information. Pour cela, la méthodologie combine les trois principales approches qui sont traditionnellement proposées pour faciliter le post-traitement des règles : indices de règle, interactivité, et représentation visuelle.

La méthodologie *RF* consiste à laisser l'utilisateur explorer par lui-même des petits ensembles successifs de règles au moyen d'une visualisation interactive des règles et de leurs indices. En d'autres termes, l'utilisateur dirige une **suite d'explorations locales visuelles** en fonction de son intérêt pour les règles. Ainsi, un ensemble volumineux de règles est exploré sous-ensemble après sous-ensemble, de telle façon que l'utilisateur n'ait jamais à l'appréhender dans sa globalité. Cette approche est fondée sur :

- les principes cognitifs de traitement de l'information de Montgomery dans le contexte des modèles de décision [Mon83] ;
- les principes de visualisation d'information de Bertin pour la construction de représentations efficaces [Ber67].

Sur la base des principes cognitifs de Montgomery, nous développons le concept de *relation de voisinage* entre règles : des relations faisant sens pour l'utilisateur qui lui permettent d'isoler des sous-ensembles de règles limités et de naviguer entre les sous-ensembles. Ces relations de voisinage constituent une originalité forte de notre méthodologie en comparaison aux autre approches d'exploration de règles. Quant aux principes de visualisation de Bertin, nous les utilisons pour mettre en valeur les indices de règle et faciliter la reconnaissance des meilleures règles dans notre méthodologie.

Dans la prochaine et dernière partie de cette thèse, nous présentons une implémentation de la méthodologie *RF* nommée *ARVis*.

Troisième partie

ARVis, un outil de visualisation pour l'extraction et l'exploration interactives des règles d'association

Visualisation interactive des règles avec *ARVis*

5

Sommaire

La méthodologie *RF* présentée dans le chapitre 4 définit des principes de base pour l'élaboration d'un outil d'exploration de règles d'association. La méthodologie peut cependant être implémentée de multiples façons. En particulier, diverses possibilités sont envisageables pour les relations de voisinage et pour l'encodage graphique. Dans ce chapitre, nous décrivons les choix qui ont été effectués pour mettre en oeuvre la méthodologie *RF* dans l'outil de visualisation *ARVis* (*Association Rule Visualization*).

Dans les sections 5.2 et 5.3, nous présentons les fonctionnalités des deux versions d'*ARVis* qui ont été réalisées. Pour cela, nous nous référons au modèle des outils de visualisation de Card, Mackinlay, et Shneiderman [CMS99], qui met

en évidence trois composantes : les transformations sur les entrées, l'encodage graphique, et les transformations sur la vue (voir chapitre 4 à la page page 92). Nous détaillons ensuite l'implémentation d'*ARVis*, et enfin décrivons quelques exemples d'utilisation.

5.1 Terminologie et notations

I est l'ensemble des items décrivant les données étudiées, et $\mathcal{R}$ est un ensemble de règles d'association extrait à partir de ces données, et décrit par des indices de règle. Sans nuire à la généralité de nos propos, nous considérons que $\mathcal{R}$ contient uniquement des règles à conclusion simple, c'est-à-dire ne comportant qu'un seul item en conclusion. Cette restriction est souvent effectuée en recherche de règles d'association (dans la définition initiale des règles d'association [AIS93], une règle ne comporte d'ailleurs qu'un seul item en conclusion). Spontanément, l'utilisateur porte en effet davantage d'intérêt à de telles règles qu'à des règles à conclusion multiple, qui sont moins intelligibles.

Dans cette partie de la thèse, nous adoptons une notation ensembliste pour les itemsets plutôt qu'une notation logique utilisant des conjonctions : un itemset $(A \wedge B \wedge C)$ est noté comme l'ensemble (A, B, C). Ceci nous permet de pouvoir effectuer des opérations ensemblistes sur les itemsets. En conséquence, une règle $(A \wedge B \wedge C) \rightarrow (D \wedge E)$ est notée $(A, B, C) \rightarrow (D, E)$.

5.2 *ARVis* version 1.1

La première version d'*ARVis* qui a été développée (présentée dans [BGB03c] et [BGB03d]) n'implémente la méthodologie *RF* que partiellement. Elle fait suite aux travaux de thèse de Lehn sur la visualisation des règles d'association par des graphes [Leh00], travaux qui ont débouché sur la réalisation de l'outil PerformanSe-FELIX.

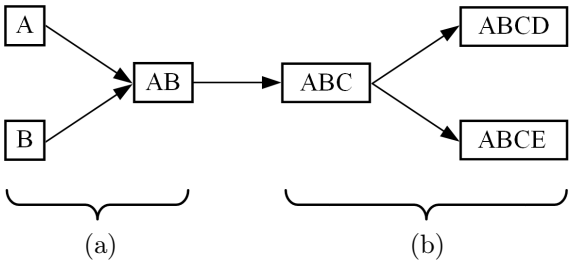


FIG. 5.1 – Exemple d'un graphe d'itemsets

5.2.1 Transformations sur les entrées

Dans PerformanSe-FELIX, l'utilisateur visualise les règles par un graphe d'itemsets (voir chapitre 4 à la page 89). Pour explorer les règles, il peut

développer le graphe à sa guise, soit en faisant apparaître les sur-itemsets d'un itemset, soit en faisant apparaître les sous-itemsets d'un itemset. Par exemple, avec un ensemble de cinq items $I = \{A, B, C, D, E\}$, à partir d'une règle $(A, B) \rightarrow (C)$ l'utilisateur peut faire apparaître les règles plus spécifiques $(A, B, C) \rightarrow (D)$ et $(A, B, C) \rightarrow (E)$ (figure 5.1.(b)), ou bien les règles plus générales $(A) \rightarrow (B)$ et $(B) \rightarrow (A)$ (figure 5.1.(a)). Ces deux approches sont reprises dans *ARVis 1.1* sous la forme des relations de voisinage suivantes.

– *Chaînage avant* :

$$\Pi_1\Big(X \rightarrow y\Big) = \Big\{X \cup y \rightarrow z \;\Big|\; z \in I \setminus (X \cup y)\Big\}$$

– *Généralisation de la prémisse* :

$$\Pi_2\Big(X \rightarrow y\Big) = \Big\{X \setminus z \rightarrow z \;\Big|\; z \in X\Big\}$$

(Dans un souci de simplicité, les relations de voisinage ne sont pas définies dans ce chapitre comme des relations binaires sur $\mathcal{R}$ mais comme des fonctions Π de $\mathcal{R}$ vers $2^{\mathcal{R}}$ qui associent à chaque règle le sous-ensemble de ses voisines. Par ailleurs, les majuscules X désignent des itemsets et les minuscules y désignent des items. Nous notons $X \cup y$ au lieu de $X \cup \{y\}$ et $X \setminus y$ au lieu de $X \setminus \{y\}$.)

Exemple. Nous reprenons l'exemple de la figure 5.1 :

$$\Pi_1\Big((A, B) \rightarrow (C)\Big) = \Big\{(A, B, C) \rightarrow (D)\ ;\ (A, B, C) \rightarrow (E)\Big\}$$

$$\Pi_2\Big((A, B) \rightarrow (C)\Big) = \Big\{(A) \rightarrow (B)\ ;\ (B) \rightarrow (A)\Big\}$$ □

La relation de voisinage Π_1 se nomme *Chaînage avant* car elle s'apparente aux inférences réalisées par un moteur d'inférence en chaînage avant (voir figure 5.2) : quand une règle $X \rightarrow y$ est activée, le fait y entre dans la base de connaissances et peut donc être utilisé avec X pour activer de nouvelles règles et inférer de nouveaux faits z. Le chaînage arrière ne peut être envisagé avec des règles à conclusion simple. Quant à la relation de voisinage *Généralisation de la prémisse*, elle est complémentaire à *Chaînage avant*. En effet, après avoir appliqué *Chaînage avant* sur une règle r, on peut retrouver r en utilisant *Généralisation de la prémisse* (figure 5.3). Il s'agit d'une réciprocité partielle entre les deux relations.

Dans *ARVis 1.1*, les relations de voisinage Π_1 et Π_2 sont utilisées conjointement au sein d'une relation de voisinage globale $\Pi = \Pi_1 \cup \Pi_2$. Ainsi, chaque sous-ensemble de règles contient deux catégories de règles :

– les règles spécifiques, issues de Π_1,
– les règles générales, issues de Π_2.

Nous verrons par la suite (section 5.2.2) que les deux types de règles sont clairement séparés dans la visualisation.

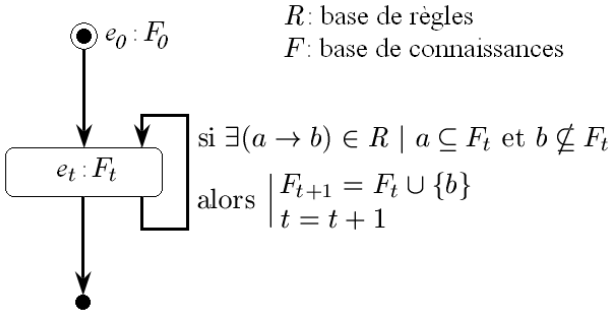


FIG. 5.2 – Moteur d'inférence en chaînage avant

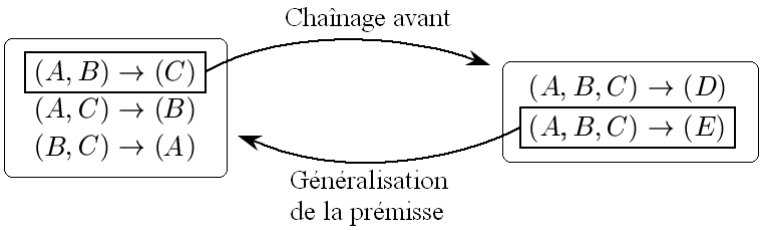


FIG. 5.3 – Exemple de réciprocité partielle des relations *Chaînage avant* et *Généralisation de la prémisse*

5.2.2 Encodage graphique

Comme nous l'avons vu en partie 2, les corpus de règles d'association sont généralement très volumineux. Afin de supporter de grandes quantités de règles tout en mettant en évidence les meilleures d'entre elles, nous avons choisi une représentation 3D pour implémenter la méthodologie *RF*. Plus précisément, la visualisation de chaque sous-ensemble de règles repose sur la métaphore du paysage d'information (*information landscape* [And95]) présentée au chapitre 4 à la page 102. Dans le paysage 3D, une règle est représentée par l'objet graphique suivant : une sphère posée au-dessus d'un cône.

Comme le préconise la méthodologie *RF* (voir section 4.5.2 page 112), nous optons pour un encodage graphique qui s'appuie sur des positions et des tailles pour mettre en valeur les indices de règle. Dans un paysage d'information, la position d'un objet pourrait nous permettre d'encoder deux indices. Cependant, étant donné que des règles différentes peuvent présenter les mêmes valeurs d'indices, il est nécessaire de laisser une dimension libre pour pouvoir répartir les objets dans l'espace et éviter qu'ils ne se chevauchent (cette technique est appelée *jitterring* lorsque les valeurs prises sur la dimension libre sont fixées aléatoirement). Un seul indice de règle est donc encodé par la position dans *ARVis*. De plus, afin de faciliter la perception de la profondeur dans l'espace 3D, les objets sont disposés sur une arène semi-circulaire[1]. Ceci revient à lier profondeur et

[1] Un choix similaire est effectué dans le gestionnaire de favoris Data Mountain de Microsoft Research, où le sol du paysage est un plan incliné [RCL+98].

hauteur : plus un objet se situe à l'arrière de la scène et plus il est placé en hauteur (figure 5.4). L'arène permet également de réduire les occultations entre objets. Au final, nous optons pour l'encodage graphique suivant (figure 5.5) :

- la position de l'objet représente l'intensité d'implication,
- la surface visible du cône représente la confiance,
- la surface visible de la sphère représente le support,
- la couleur de l'objet représente une moyenne pondérée de la confiance et de l'intensité d'implication, ce qui donne une idée synthétique de la qualité de la règle.

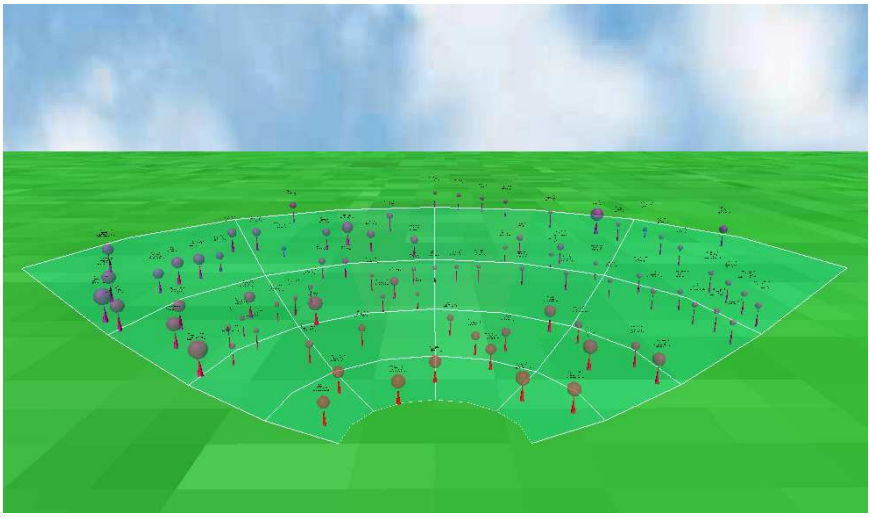
L'encodage graphique met fortement en évidence les règles de bonne qualité, dont la visualisation et l'accès sont facilités par rapport aux règles plus mauvaises. Plus précisément, une grande sphère rouge posée sur un haut cône à l'avant de l'arène (sur les marches basses) représente une règle dont le support, la confiance et l'intensité d'implication sont élevés, tandis qu'une petite sphère bleue posée sur un cône bas au fond de l'arène (sur les marches hautes) représente une règle dont les trois mesures sont faibles. En outre, des étiquettes textuelles complémentaires sont affichées au-dessus de chaque objet pour donner le nom de la règle correspondante. Elles indiquent également les valeurs numériques du support, de la confiance et de l'intensité d'implication.

Dans *ARVis 1.1*, étant donné que la relation de voisinage combine spécialisation et généralisation, chaque paysage 3D contient deux arènes : une pour les règles spécifiques et une pour les règles générales. Afin de bien dissocier les deux types de règles, les deux arènes sont placées face à face et l'utilisateur débute l'exploration de chaque paysage entre les deux arènes. De plus, le paysage que l'utilisateur peut faire apparaître en appliquant une relation de voisinage sur une règle est affiché en version rétrécie dans la sphère correspondante (figure 5.5). Ces objets miniatures visibles par transparence assistent l'utilisateur dans sa navigation en lui permettant par exemple d'anticiper qu'un sous-ensemble contient des règles de bonne qualité, ou au contraire ne contient aucune règle[2].

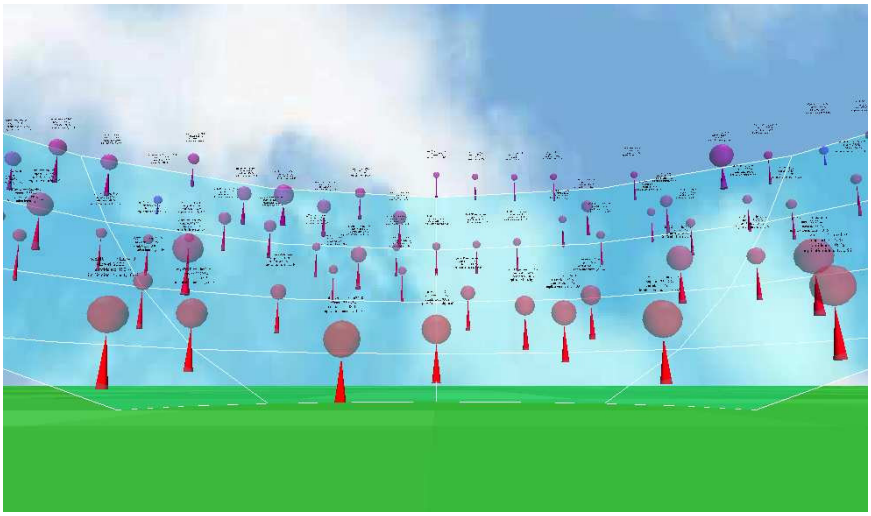
5.2.3 Transformations sur la vue

Au début de la visite d'un paysage 3D, l'utilisateur est placé au centre des deux arènes, et bénéficie ainsi d'un point de vue global sur les règles d'une arène. Lors de cette vision d'ensemble, il est aisé de localiser les meilleures règles. L'utilisateur peut ensuite visiter librement le paysage pour examiner les règles de plus près en modifiant le point de vue sur la scène.

[2]Dans un souci d'intelligibilité de l'affichage, les objets représentant des règles spécifiques ne contiennent en version rétrécie que des règles spécifiques, tandis que les objets des règles générales ne contiennent que des règles générales. L'anticipation sur le prochain paysage à afficher ne porte donc que sur une seule des deux arènes.



(a)



(b)

(c)

FIG. 5.4 – Un paysage de règles dans *ARVis*

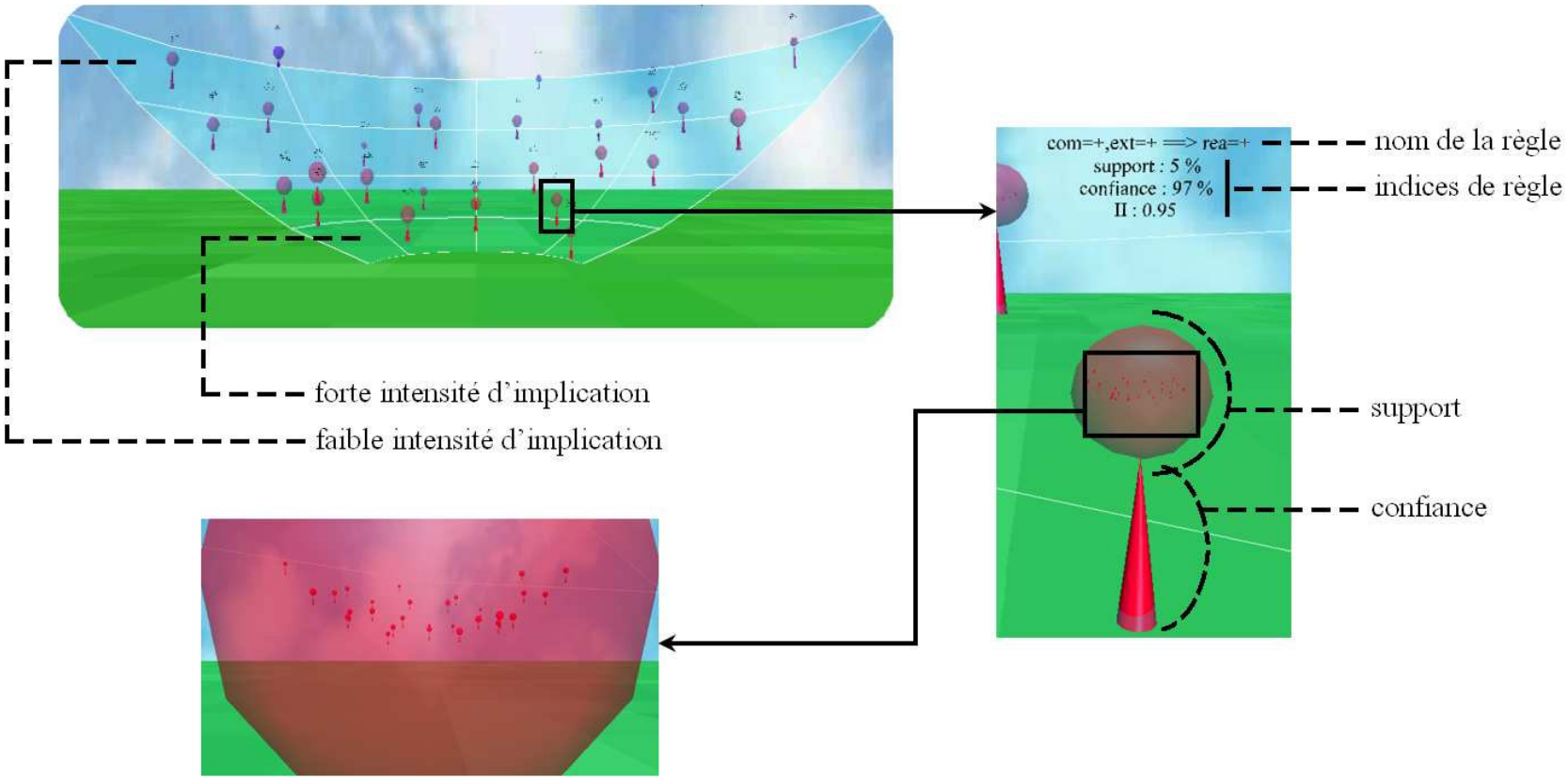


FIG. 5.5 – Encodage graphique dans *ARVis 1.1*

5.3 *ARVis* version 1.2

ARVis 1.2 est l'implémentation la plus complète de la méthodologie *RF*. Contrairement à la version *1.1*, elle propose de multiples relations de voisinage et utilise un indice descriptif d'écart à l'équilibre, un indice descriptif d'écart à l'indépendance, un indice statistique d'écart à l'équilibre, et un indice statistique d'écart à l'indépendance pour évaluer les règles (ce que la méthodologie *RF* préconise, voir section 4.5.2 page 112).

5.3.1 Transformations sur les entrées

Huit relations de voisinage sont proposées dans *ARVis 1.2*, la plupart d'entre elles étant des relations de spécialisation ou de généralisation. Spécialisation et généralisation sont en effet les deux processus cognitifs fondamentaux pour la génération de nouvelles règles (voir l'étude des processus de raisonnement dans [HHNT86]). Les relations de voisinage dans *ARVis 1.2* ont également la particularité d'intégrer des indices de règle :

- le support (indice de similarité),
  noté sp,
- la confiance (indice descriptif d'écart à l'équilibre),
  noté cf,
- le lift (indice descriptif d'écart à l'indépendance),
  noté li,
- l'intensité d'implication (indice statistique d'écart à l'indépendance),
  noté ii,
- IPEE (indice statistique d'écart à l'équilibre),
  noté ip.

Chaque indice est associé à un seuil minimal et un seuil maximal fixés par l'utilisateur afin de filtrer les règles :

$$min_{sp},\ min_{cf},\ min_{li},\ min_{ii},\ \text{et}\ min_{ip}$$

$$max_{sp},\ max_{cf},\ max_{li},\ max_{ii},\ \text{et}\ max_{ip}$$

Les seuils peuvent être modifiés par l'utilisateur à tout instant pendant la navigation dans $\mathcal{R}$. Pour définir les relations de voisinage, nous regroupons tous les seuils au sein de la fonction booléenne *BonneQualité* :

$$\forall\ r \in \mathcal{R}, BonneQualité(r) \Leftrightarrow \left(\begin{array}{ll} min_{sp} \leq sp(r) \leq max_{sp} & \text{et} \\ min_{cf} \leq cf(r) \leq max_{cf} & \text{et} \\ min_{li} \leq li(r) \leq max_{li} & \text{et} \\ min_{ii} \leq ii(r) \leq max_{ii} & \text{et} \\ min_{ip} \leq ip(r) \leq max_{ip} & \end{array} \right)$$

Bien que la plupart des outils de règles d'association n'en proposent pas, les seuils maximaux facilitent la tâche de l'utilisateur. Par exemple, les règles

de très fort support et très forte confiance sont souvent déjà connues des utilisateurs ; les supprimer permet de mettre en évidence des règles plus intéressantes.

◇ *Relations de spécialisation*

– *Spécialisation concordante* :

$$\Pi_1\Big(X \to y\Big) = \Big\{X \cup z \to y \;\Big|\; z \in I \backslash (X \cup y) \wedge \mathit{BonneQualité}(X \cup z \to y)\Big\}$$

– *Spécialisation d'exception* :

$$\Pi_2\Big(X \to y\Big) = \Big\{X \cup z \to \overline{y} \;\Big|\; z \in I \backslash (X \cup \overline{y}) \wedge \mathit{BonneQualité}(X \cup z \to \overline{y})\Big\}$$ [3]

– *Chaînage avant* :

$$\Pi_3\Big(X \to y\Big) = \Big\{X \cup y \to z \;\Big|\; z \in I \backslash (X \cup y) \wedge \mathit{BonneQualité}(X \cup y \to z)\Big\}$$

Exemples. L'ensemble des items est $I = \{A, B, C, D, E\}$. Ici, nous faisons abstraction de la fonction *BonneQualité* pour ne considérer que la syntaxe des règles.

$$\Pi_1\Big((A,B) \to (C)\Big) = \Big\{(A,B,D) \to (C)\ ;\ (A,B,E) \to (C)\Big\}$$

$$\Pi_2\Big((A,B) \to (C)\Big) = \Big\{(A,B,D) \to (\overline{C})\ ;\ (A,B,E) \to (\overline{C})\Big\}$$

$$\Pi_3\Big((A,B) \to (C)\Big) = \Big\{(A,B,C) \to (D)\ ;\ (A,B,C) \to (E)\Big\}$$ □

Dans [HHNT86], Holland *et al.* soulignent qu'une règle trop générale peut être spécialisée en deux types de règles complémentaires : les règles exceptions et les règles concordantes. Les règles exceptions visent à expliquer les contre-exemples de la règle générale, tandis que les règles concordantes visent à mieux expliquer les exemples. Par exemple, une règle "Si X est un chien alors X est gentil" peut être spécialisée en deux règles "Si X est un chien et X est muselé alors X est méchant" et "Si X un chien et X n'est pas muselé alors X est gentil". L'intérêt des règles exceptions en ECD a été largement confirmé (voir par exemple [HLSL00] et [SZ05]). Sur la base de ces deux types de spécialisation, nous proposons les relations de voisinage *Spécialisation concordante* et *Spécialisation d'exception*. La relation *Chaînage avant* est quant à elle reprise de *ARVis 1.1* (voir description section 5.2.1).

[3] $\overline{y}$ désigne ici n'importe quel item provenant de la même variable que y mais présentant une modalité différente. Par exemple, si y est l'item *couleur_des_yeux=bleu*, alors $\overline{y}$ peut être *couleur_des_yeux=marron* ou *couleur_des_yeux=vert*.

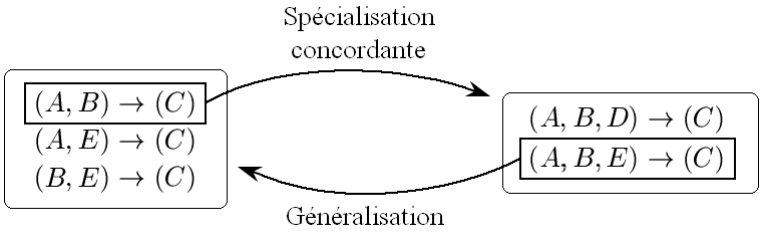


FIG. 5.6 – Réciprocité partielle de *Spécialisation concordante* et *Généralisation* (pour un ensemble de cinq items $I = \{A, B, C, D, E\}$)

◇ *Relations de généralisation*

– *Généralisation* :

$$\Pi_4\Big(X \rightarrow y\Big) = \Big\{X \setminus z \rightarrow y \;\Big|\; z \in X \;\wedge\; \mathit{BonneQualité}(X \setminus z \rightarrow y)\Big\}$$

– *Généralisation de la prémisse* :

$$\Pi_5\Big(X \rightarrow y\Big) = \Big\{X \setminus z \rightarrow z \;\Big|\; z \in X \;\wedge\; \mathit{BonneQualité}(X \setminus z \rightarrow z)\Big\}$$

Exemples.

$$\Pi_4\Big((A, B, C) \rightarrow (D)\Big) = \Big\{(A, B) \rightarrow (D)\ ;\ (A, C) \rightarrow (D)\ ;\ (B, C) \rightarrow (D)\Big\}$$

$$\Pi_5\Big((A, B, C) \rightarrow (D)\Big) = \Big\{(A, B) \rightarrow (C)\ ;\ (A, C) \rightarrow (B)\ ;\ (B, C) \rightarrow (A)\Big\}\ \square$$

La relation *Généralisation* consiste à simplifier la prémisse d'une règle (processus de simplification des conditions décrit dans [HHNT86]). Elle est complémentaire à *Spécialisation concordante*, puisque après avoir appliqué *Spécialisation concordante* sur une règle r, on peut retrouver r en utilisant *Généralisation* (figure 5.6). La relation *Généralisation de la prémisse* est quant à elle reprise de *ARVis 1.1* (voir description section 5.2.1).

◇ *Autres relations*

– *Prémisse commune* :

$$\Pi_6\Big(X \rightarrow y\Big) = \Big\{X \rightarrow z \;\Big|\; z \in I \setminus X \;\wedge\; \mathit{BonneQualité}(X \rightarrow z)\Big\}$$

– *Conclusion commune* :

$$\Pi_7\Big(X \rightarrow y\Big) = \Big\{z \rightarrow y \;\Big|\; z \in I \setminus y \;\wedge\; \mathit{BonneQualité}(z \rightarrow y)\Big\}$$

– *Items communs* :

$$\Pi_8\Big(X \rightarrow y\Big) = \Big\{(X \cup y) \backslash z \rightarrow z \;\Big|\; z \in X \cup y \wedge \mathit{BonneQualité}((X \cup y) \backslash z \rightarrow z)\Big\}$$

Exemples.

$$\Pi_6\Big((A,B) \rightarrow (C)\Big) = \Big\{(A,B) \rightarrow (C)\ ;\ (A,B) \rightarrow (D)\ ;\ (A,B) \rightarrow (E)\Big\}$$

$$\Pi_7\Big((A,B) \rightarrow (C)\Big) = \Big\{(A) \rightarrow (C)\ ;\ (B) \rightarrow (C)\ ;\ (D) \rightarrow (C)\ ;\ (E) \rightarrow (C)\Big\}$$

$$\Pi_8\Big((A,B) \rightarrow (C)\Big) = \Big\{(A,B) \rightarrow (C)\ ;\ (A,C) \rightarrow (B)\ ;\ (B,C) \rightarrow (A)\Big\} \quad \square$$

Les relations *Prémisse commune* et *Conclusion commune* préservent la prémisse et changent la conclusion, ou vice versa. *Items communs* permet de permuter les items dans une règle. Toutes les règles produites par cette relation sont vérifiées par la même population d'individus dans les données.

Imaginons que l'utilisateur applique une relation de voisinage Π sur une règle r. Ceci génère et affiche un nouveau sous-ensemble $S = \Pi(r)$ contenant toutes les règles voisines de r selon Π. Nous appelons r la *règle de transition*, car c'est elle qui permet le passage d'un sous-ensemble à un autre. En fonction de la relation Π choisie, S peut contenir ou ne pas contenir la règle de transition (les relations de voisinage ne sont pas nécessairement réflexives –voir chapitre 4). Dans *ARVis 1.2*, nous ajoutons systématiquement la règle de transition à tout sous-ensemble généré par une relation de voisinage. Ceci permet de pouvoir effectuer des comparaisons entre la règle de transition et ses règles voisines. Par exemple, avec la relation de voisinage *Généralisation*, il est intéressant de comparer une règle à ses voisines afin de repérer les items superflus dans la règle (ceux dont la suppression ne dégrade pas la qualité de la règle). Réciproquement, avec la relation *Spécialisation concordante*, comparer une règle à ses voisines permet de vérifier si l'ajout d'un nouvel item en prémisse améliore ou non la prédiction de la conclusion.

5.3.2 Encodage graphique

ARVis 1.2 reprend la métaphore du paysage d'information de la version *1.1*, avec les mêmes objets graphiques : une sphère posée au-dessus d'un cône. Ces derniers sont également disposés sur une arène afin de faciliter la perception de la profondeur dans l'espace 3D et de réduire les occultations entre objets. Chaque paysage 3D ne contient par contre qu'une seule arène. En effet, les relations de voisinage dans *ARVis 1.2* étant plus nombreuses, elle sont utilisées indépendamment les unes des autres et non pas conjointement comme dans *ARVis 1.1*.

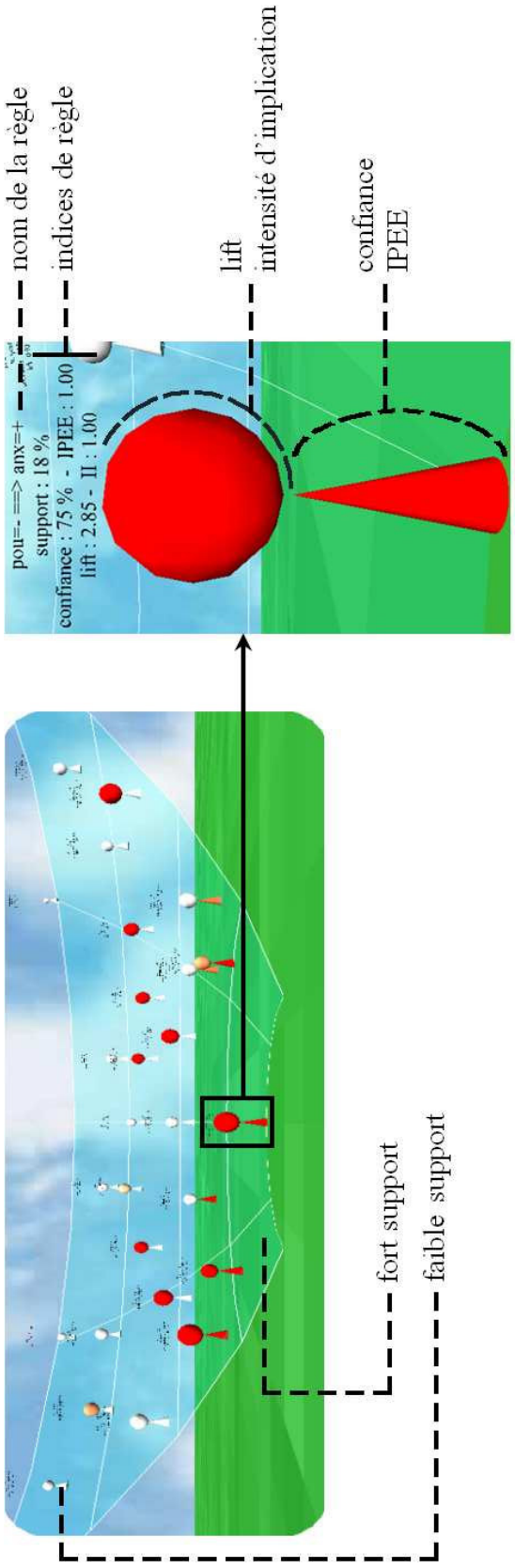


FIG. 5.7 – Encodage graphique dans *ARVis 1.2*

L'encodage graphique de *ARVis 1.2* diffère de celui de la version *1.1* puisqu'il intègre les cinq indices de règle précédemment énoncés : support, confiance, lift, intensité d'implication, et IPEE. Comme le préconise la méthodologie *RF* (voir section 4.5.2 page 112), nous optons pour un encodage graphique qui s'appuie sur des positions et des tailles pour mettre en valeur les indices de règle (figure 5.7) :

- la position de l'objet représente le support,
- la surface visible du cône représente la confiance,
- la surface visible de la sphère représente le lift (dans un souci d'homogénéité vis-à-vis des autres indices, le lift est normalisé entre 0 et 1),
- la luminosité de la sphère représente l'intensité d'implication,
- la luminosité du cône représente IPEE.

Ainsi, **la sphère est dédiée aux indices d'écart à l'indépendance**, tandis que **le cône est dédié aux indices d'écart à l'équilibre**. L'encodage rend clairement identifiables les règles de bonne qualité, dont la visualisation et l'accès sont facilités par rapport aux règles plus mauvaises. Les sphères ne contiennent pas d'objets miniatures permettant d'anticiper la navigation, puisque *ARVis 1.2* propose huit relations de voisinage et qu'il n'est pas possible de prévoir celle que l'utilisateur va appliquer. Par ailleurs, dans chaque paysage, la règle de transition est représentée par un objet semi-transparent pour être distinguée des autres règles et ainsi permettre les comparaisons.

Cet encodage graphique ne fait pas appel à la couleur, contrairement à l'encodage adopté dans *ARVis 1.1*. En effet, comme nous l'avons vu au chapitre 4, la couleur est adaptée à l'encodage de variables nominales mais elle est généralement déconseillée pour l'encodage de variables ordinales ou quantitatives (voir tableau 4.2 page 95). Si nous avons tout de même décidé d'utiliser la couleur dans *ARVis 1.1*, c'est parce que l'information correspondante y est redondante (moyenne des indices de règle), ce qui en diminue le caractère stratégique. Pour représenter l'intensité d'implication et IPEE dans *ARVis 1.2*, nous employons, à la place de la couleur, la luminosité[4] qui, elle, est adaptée à l'encodage de variables ordinales. Pour cela, les deux indices de règle sont discrétisés en cinq modalités (il est conseillé de ne pas représenter plus de six ou sept modalités avec la luminosité comme avec la couleur [Ber67] [Wil05]). Comme dans *ARVis 1.1*, les valeurs précises des indices sont indiquées sur les étiquettes textuelles au-dessus des objets.

5.3.3 Transformations sur la vue

ARVis 1.2 reprend les transformations sur la vue de la version *1.1* : l'utilisateur peut visiter librement le paysage 3D en modifiant le point de vue sur la scène. Au début de la visite d'un paysage, il est placé devant l'arène, et bénéficie ainsi d'un point de vue global sur les règles. Par rapport à *ARVis 1.1*, des aides à la visite des paysages sont ajoutées sous la forme de points de vue prédéfinis dans le paysage, donnant une vision globale de l'arène (comme ceux montrés

[4] La couleur constante choisie est le rouge. Les objets peuvent donc varier du blanc au noir en passant par des teintes rouges. Cependant pour que les étiquettes textuelles (écrites en noir) restent visibles, nous faisons varier les objets du blanc jusqu'à un rouge saturé, sans aller jusqu'au noir.

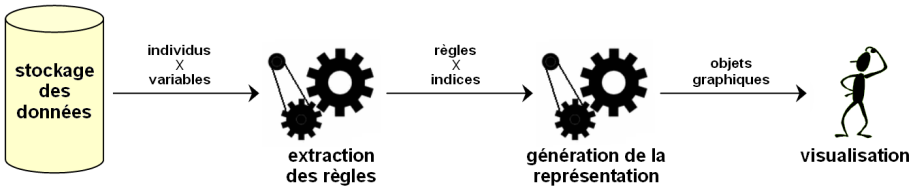


FIG. 5.8 – Architecture logique d'*ARVis*

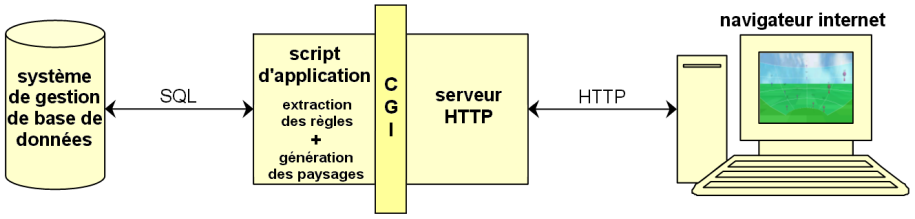


FIG. 5.9 – Architecture physique d'*ARVis*

dans la figure 5.4 page 122) ou une vision rapprochée de chaque objet. Un menu listant toutes les règles permet d'afficher directement le point de vue rapproché sur l'objet correspondant. Ainsi, *ARVis 1.2* permet à l'utilisateur de trouver les règles qui l'intéressent que ses critères de recherche soient fondés sur les indices de règle (utilisation de la représentation visuelle) ou sur les items (utilisation du menu).

5.4 Implémentation

5.4.1 Architecture

ARVis 1.1 et *1.2* sont construits sur une architecture client-serveur trois-tiers (figures 5.8 et 5.9) :

- le serveur de données est le système de gestion de bases de données relationnelles PostgreSQL[5],
- la couche applicative est un script écrit en Perl et exécuté par un serveur HTTP *via* une passerelle CGI[6],
- le client est un navigateur internet.

C'est sur le client que s'effectue la visualisation des sous-ensembles de règles par l'utilisateur (affichage et visite des paysages 3D). Le script Perl est quant à lui composé de deux modules.

- Un premier module prend en charge l'extraction des règles (et des valeurs

[5] www.postgresql.org

[6] *Common Gateway Interface* est un protocole standard permettant d'interfacer un serveur HTTP avec des applications.

d'indices) en implémentant des algorithmes à contraintes adaptés aux relations de voisinage. Ce module est interactif : il produit les sous-ensembles de règles à la demande, au fur et à mesure de la navigation de l'utilisateur (aucune règle n'est produite à l'avance).
- Un second module prend en charge la génération des paysages 3D. Il s'agit également d'un module interactif qui génère les paysages au fur et à mesure de la navigation de l'utilisateur (aucun paysage n'est généré à l'avance).

Contrairement à l'extraction de règles, la génération des paysages ne nécessite aucun accès à la base de données et est très peu consommatrice en temps.

Le chapitre 6 est entièrement dédié au module d'extraction. Ci-dessous, nous décrivons le module de génération des paysages ainsi que l'implémentation de la visualisation sur le client.

5.4.2 Génération des paysages sur le serveur

Un monde virtuel en VRML

ARVis génère les paysages 3D en VRML (*Virtual Reality Modeling Language*). Il s'agit d'un langage de référence (norme ISO) pour la description de mondes virtuels et interactifs en 3D, créé en 1995 et promu par le *Web3D Consortium*. A l'origine, VRML était destiné à devenir le standard de diffusion de scènes 3D sur Internet. Ceci ne s'est pas réalisé puisque le langage n'a jamais réussi à s'imposer face aux développements propriétaires. Cependant, VRML n'en reste pas moins aujourd'hui le format universel pour les contenus 3D, implémenté par tous les logiciels de création et d'édition 3D. La dernière version du langage (dialecte XML nommé X3D ou *eXtensible 3D*, normalisé en 2004) est évolutive, c'est-à-dire qu'elle accepte l'ajout de nouvelles fonctionnalités. Elle est intégrée à la norme audio/vidéo MPEG-4, qui supporte ainsi les contenus 3D.

Les mondes virtuels écrits en VRML peuvent être affichés et explorés au moyen d'outils de visualisation appelés navigateurs VRML. Ces outils disposent de primitives de navigation standards (marcher, voler, examiner, etc.), activables à la souris ou au clavier, qui modifient le point de vue de l'utilisateur sur le monde virtuel. Dans *ARVis*, le navigateur VRML est un module d'extension (*plug-in*) qui équipe le navigateur internet sur le poste client.

Pour chaque paysage à visualiser, *ARVis* génère le code VRML :
- du sol,
- du ciel,
- d'une source de lumière,
- de l'arène (ou des arènes pour *ARVis 1.1*)
- des objets graphiques (cône et sphère) accompagnés de leurs étiquettes textuelles,
- dans *ARVis 1.2*, des points de vue prédéfinis pour la vision globale de l'arène et pour la vision rapprochée de chaque objet.

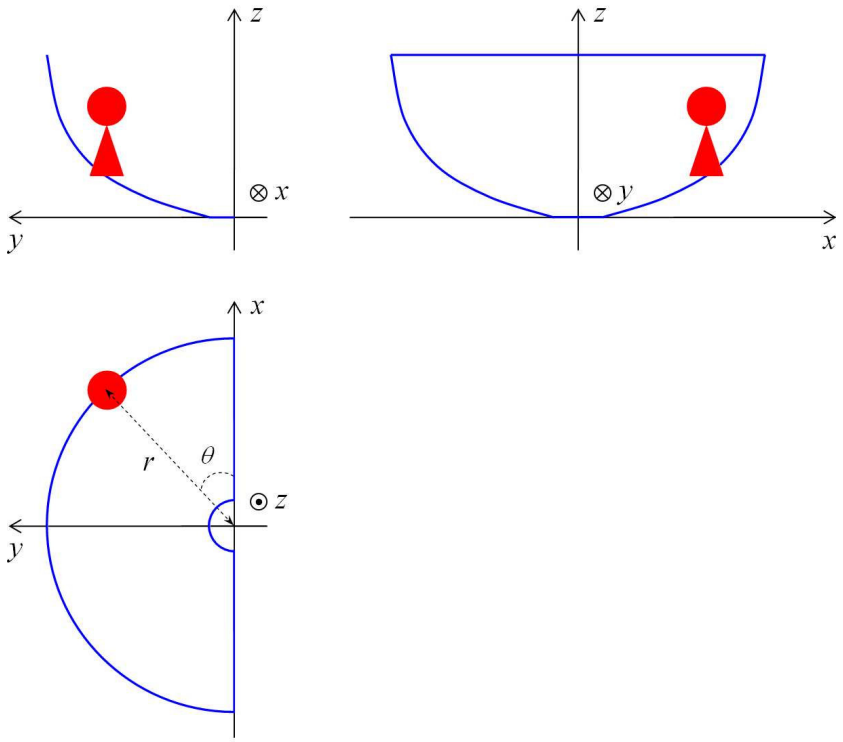


FIG. 5.10 – Le placement des objets est déterminé en coordonnées cylindriques (r, θ, z).

Création des objets et arènes

La taille d'un objet moyen dans un monde virtuel VRML doit être de l'ordre d'une unité de longueur VRML[7]. Ainsi, l'encodage graphique d'*ARVis* est traduit de la manière suivante en VRML :

- pour la sphère, $rayon = \sqrt{support}$ dans *ARVis 1.1*, et $rayon = \sqrt{lift}$ dans *ARVis 1.2*,
- pour le cône, $hauteur = 2 \times confiance$ et $rayon = 0.4$,
- pour la position, l'algorithme de placement (algorithme 5.1) fournit les coordonnées du point de référence qu'est le centre de la base du cône (voir figure 5.10).

L'arène est le morceau de parabole d'équation $(z = \frac{r^2-100}{80}, r \in [10; 40], \theta \in [0; \pi])$ en coordonnées cylindriques. Les coordonnées (r, θ, z) retournées par l'algorithme de placement correspondent à des points de cette parabole. Les valeurs

[7] On considère généralement qu'une unité de longueur VRML dans un monde virtuel équivaut à un mètre dans le monde réel. Cette convention facilite le partage des objets virtuels dans la communauté des programmeurs VRML en garantissant une cohérence entre les tailles des objets (par exemple on peut modéliser une maison en VRML et utiliser une bibliothèque d'objets VRML pour la meubler sans risquer qu'une armoire soit aussi grande que la maison). Toutefois, elle ne convient qu'à la représentation du monde réel à taille humaine (majorité des applications du VRML).

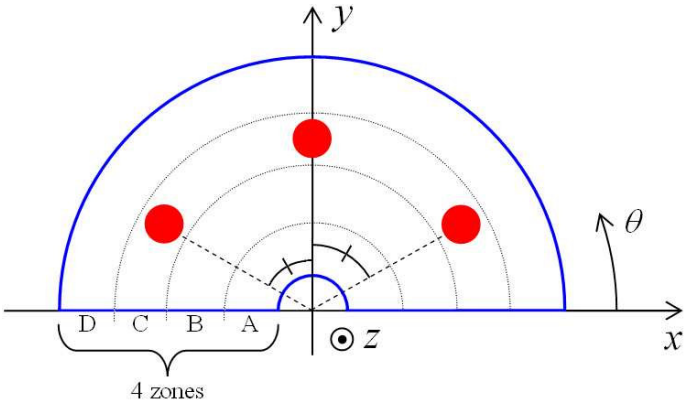


FIG. 5.11 – L'étalement des objets est maximisé dans chaque zone.

de r et de z sont déterminées à partir de l'indice de règle associé à la position dans l'encodage graphique (lignes 6 et 7 de l'algorithme 5.1). Les valeurs de θ en revanche sont calculées de manière à maximiser l'étalement des objets sur l'arène. En effet, l'arène est divisée du bas vers le haut en quatre zones, et dans chaque zone les objets sont répartis régulièrement selon des valeurs de θ équidistantes (lignes 8 à 19 de l'algorithme 5.1, voir figure 5.11). Pour les cas où les indices de règle varient faiblement sur le sous-ensemble de règles à visualiser, *ARVis* intègre une procédure optionnelle de normalisation permettant de répartir les indices sur des plages de valeurs plus larges. Ceci permet également d'éviter que les faibles valeurs d'indices conduisent à des objets trop petits pour être visibles.

La taille de l'arène est identique quel que soit le paysage considéré. De même, les tailles maximales et minimales des objets sont communes à tous les paysages. Il aurait pourtant été possible d'adapter la taille de l'arène ou la taille des objets en fonction du nombre d'objets à afficher dans le paysage. Ceci aurait favorisé la visualisation des petits sous-ensembles de règles, puisqu'un petit nombre d'objets aurait autant "rempli" l'écran qu'un grand nombre d'objets. Nous n'avons pas retenu ce choix dans *ARVis* afin de préserver les références de l'utilisateur en termes de position et de taille : dans tous les paysages, les qualités maximales et minimales correspondent toujours aux mêmes positions dans l'arène, et aux mêmes tailles des objets. Nous pensons que ces invariants visuels favorisent l'intelligibilité des paysages de règles. Pour la même raison, les points de vue prédéfinis pour la vision rapprochée des règles sont systématiquement placés à la même distance des objets, que ceux-ci soient petits ou grands.

5.4.3 Visualisation des paysages sur le client

L'utilisateur visualise et explore les paysages avec le navigateur VRML du client. Nous utilisons principalement Cortona[8] (voir interface figure 5.12), mais de nombreux produits équivalents sont disponibles.

[8] www.parallelgraphics.com

Entrées : R, ensemble de règles à représenter,
(l'indice à encoder avec la position est noté i)

Sorties : P, ensemble de triplets (r, θ, z) désignant les points de référence pour le placement des objets dans le paysage (en coordonnées cylindriques)

$r, \theta, z = 0$;
$P = \emptyset$;
$compteurZoneA, compteurZoneB = 0$;
$compteurZoneC, compteurZoneD = 0$;

pour chaque $règle \in R$ **faire**
  $r = 30 \times (1 - règle.i) + 10$;
  $z = (r^2 - 100) \div 80$;
  si $règle.i > 0.75$ **alors**
    $\theta = \pi \times \frac{compteurZoneA + 0.5}{nbreTotalObjetsZoneA}$;
    $compteurZoneA{+}{+}$;
  sinon si $règle.i > 0.5$ **alors**
    $\theta = \pi \times \frac{compteurZoneB + 0.5}{nbreTotalObjetsZoneB}$;
    $compteurZoneB{+}{+}$;
  sinon si $règle.i > 0.25$ **alors**
    $\theta = \pi \times \frac{compteurZoneC + 0.5}{nbreTotalObjetsZoneC}$;
    $compteurZoneC{+}{+}$;
  sinon
    $\theta = \pi \times \frac{compteurZoneD + 0.5}{nbreTotalObjetsZoneD}$;
    $compteurZoneD{+}{+}$;
  $P = P \cup (r, \theta, z)$;
retourne P ;

Algorithme 5.1: Algorithme de placement des objets dans *ARVis*

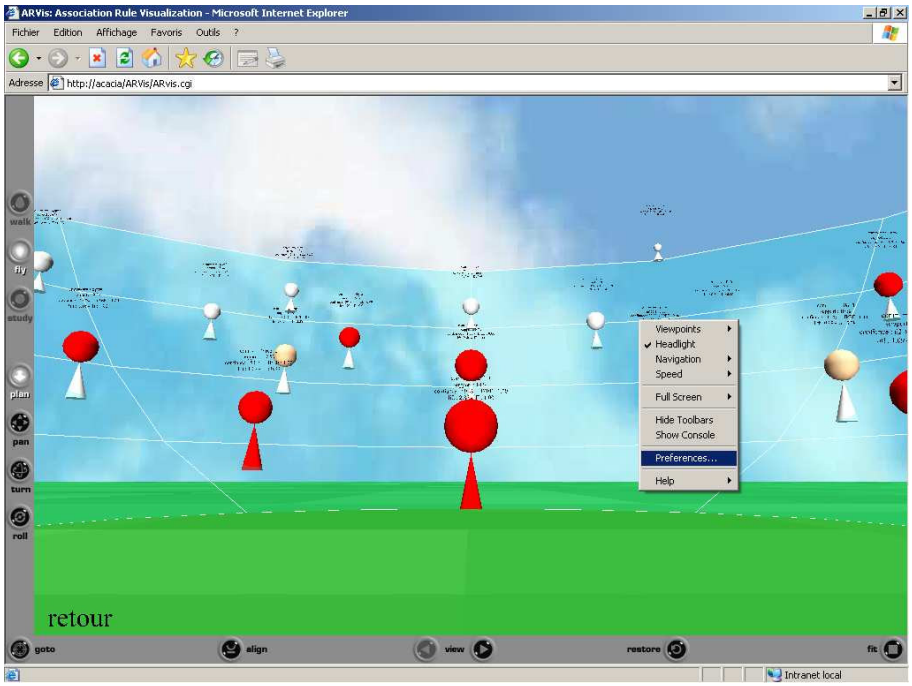


FIG. 5.12 – Interface du navigateur VRML Cortona

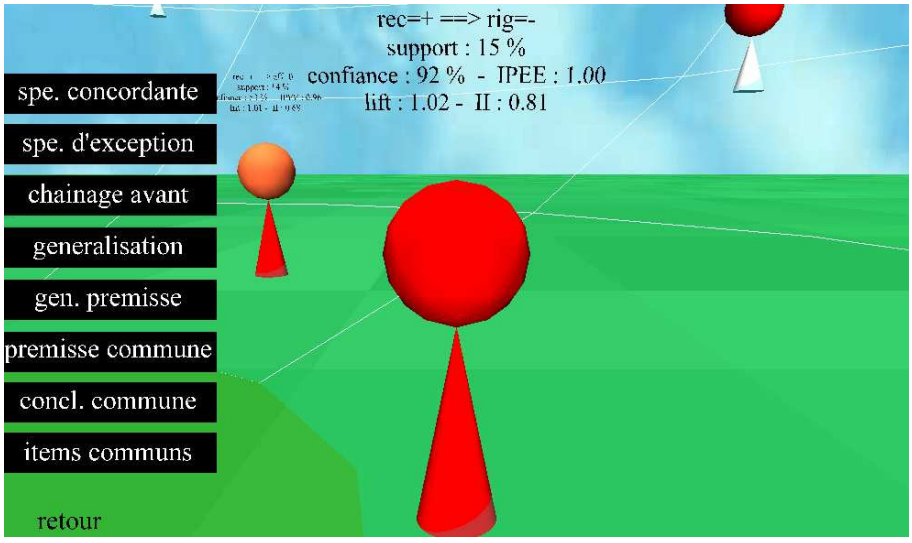


FIG. 5.13 – Menu proposant les huit relations de voisinage

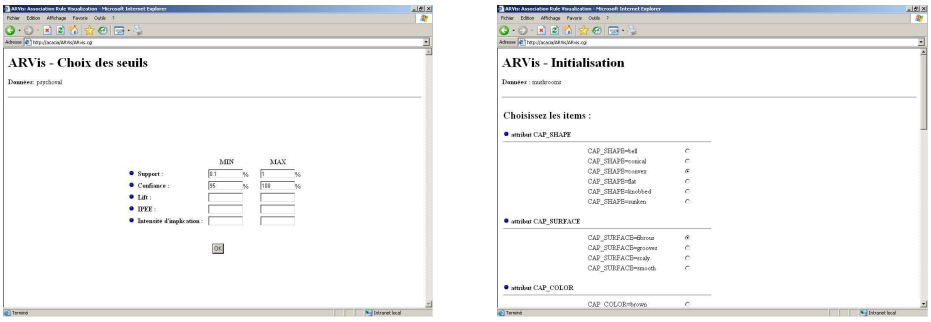


FIG. 5.14 – Interface d'initialisation

Trois types d'interactions sont possibles dans *ARVis* :

- L'utilisateur peut explorer les paysages de règles soit en se promenant librement à l'aide des primitives de déplacement du navigateur VRML, soit en utilisant les points de vue prédéfinis (*ARVis 1.2*). Il suffit de cliquer sur la sphère d'un objet pour afficher le point de vue prédéfini pour la vision rapprochée de cet objet. Concrètement, ceci revient à se déplacer automatiquement dans le paysage jusqu'à se retrouver face à l'objet.

- L'utilisateur peut filtrer les règles en modifiant les seuils minimaux et maximaux sur les indices de règle au moyen d'une interface HTML séparée. Seules les règles qui respectent les seuils sont affichées. Ceci a pour effet de faire apparaître et disparaître des objets dans le paysage.

- L'utilisateur peut naviguer d'un sous-ensemble de règles à un autre en appliquant les relations de voisinage. Dans *ARVis 1.1*, il suffit de cliquer sur le cône d'un objet pour enclencher la relation de voisinage sur la règle correspondante. Dans *ARVis 1.2*, le clique sur le cône d'un objet affiche un menu proposant les huit relations de voisinage (figure 5.5.1). En appliquant une relation de voisinage, le sous-ensemble de règles courant est remplacé par un nouveau sous-ensemble généré par le module d'extraction d'*ARVis* (voir chapitre 6). Visuellement, le paysage courant est remplacé par un nouveau paysage, ce qui donne l'impression de se déplacer virtuellement au sein de l'ensemble total des règles. Une fonction d'annulation permet de revenir aux paysages précédemment visités.

En outre, pour débuter ou recommencer une exploration de règles, l'utilisateur choisit un premier sous-ensemble de règles à visiter à l'aide d'une interface d'initialisation qui consiste en une série de pages HTML dynamiques générées par le script Perl (figure 5.14). Cette interface permet de construire l'itemset de son choix et d'afficher le paysage des règles qui contiennent cet itemset en prémisse, ou en conclusion, ou bien globalement dans les deux. L'interface permet également d'indiquer la base de données et la table de données à étudier.

Traits comportementaux			Représentation
introversion	$\neq$	extraversion	$ext \in \{-, 0, +\}$
détente	$\neq$	anxiété	$anx \in \{-, 0, +\}$
remise en cause	$\neq$	affirmation	$aff \in \{-, 0, +\}$
détermination	$\neq$	réceptivité	$rec \in \{-, 0, +\}$
improvisation	$\neq$	rigueur	$rig \in \{-, 0, +\}$
conformisme intellectuel	$\neq$	dynamisme intellectuel	$din \in \{-, 0, +\}$
conciliation	$\neq$	combativité	$com \in \{-, 0, +\}$
motivation de facilitation	$\neq$	motivation de réalisation	$rea \in \{-, 0, +\}$
motivation d'indépendance	$\neq$	motivation d'appartenance	$app \in \{-, 0, +\}$
motivation de protection	$\neq$	motivation de pouvoir	$pou \in \{-, 0, +\}$

Tab. 5.1 – Traits comportementaux

5.5 Exemples d'utilisation

Les exemples d'utilisation présentés ci-dessous ont été réalisés avec *ARVis 1.2*, l'implémentation la plus complète de la méthodologie *RF*. Les données étudiées proviennent d'une base de données de profils psychologiques appartenant à la société *PerformanSe SA*. Cette société édite des logiciels d'aide à la décision pour la gestion des ressources humaines, dédiés à l'évaluation des comportements et motivations en milieu professionnel. La base de données de profils psychologiques est utilisée pour étalonner les logiciels. Elle contient les profils de 4065 personnes (population de référence constituée d'adultes français) décrits par dix traits comportementaux bipolaires (tableau 5.1). Chacun d'eux est codé par une variable qualitative possédant trois modalités : +, 0, et – (trait fortement, moyennement, ou faiblement affirmé).

Les figures illustrant les exemples sont soit des vues globales d'une arène, soit des vues rapprochées d'une ou deux règles. Les règles pour lesquelles nous montrons une vue rapprochée sont marquées d'une croix dans les vues globales.

5.5.1 Exemple 1

L'exemple décrit ci-dessous est constitué d'une phase de spécialisation (pendant laquelle on ajoute des items pour mettre en évidence des combinaisons intéressantes) suivie d'une phase de généralisation (pendant laquelle on supprime les items superflus). Les explorations de règles que nous réalisons avec *ARVis* sont couramment constituées de plusieurs phases de spécialisation et de généralisation qui se succèdent.

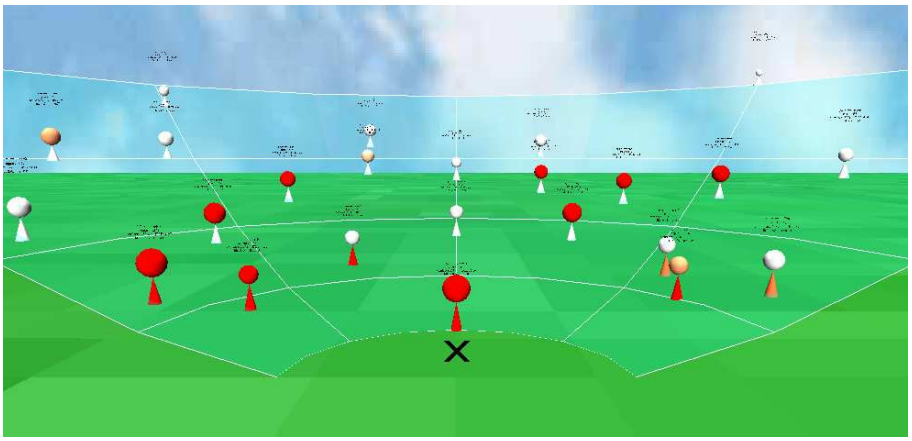


FIG. 5.15.a

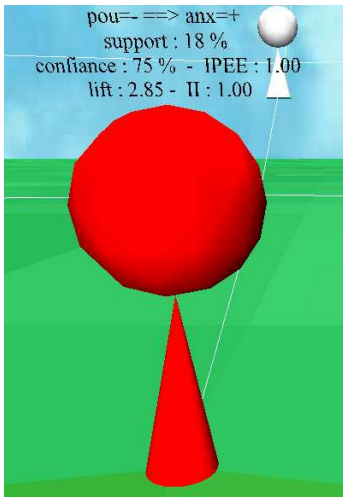


FIG. 5.15.b

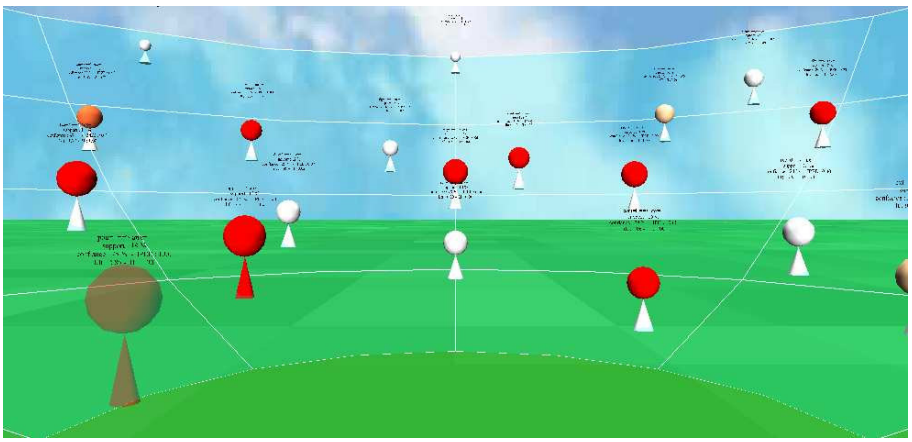


FIG. 5.15.c

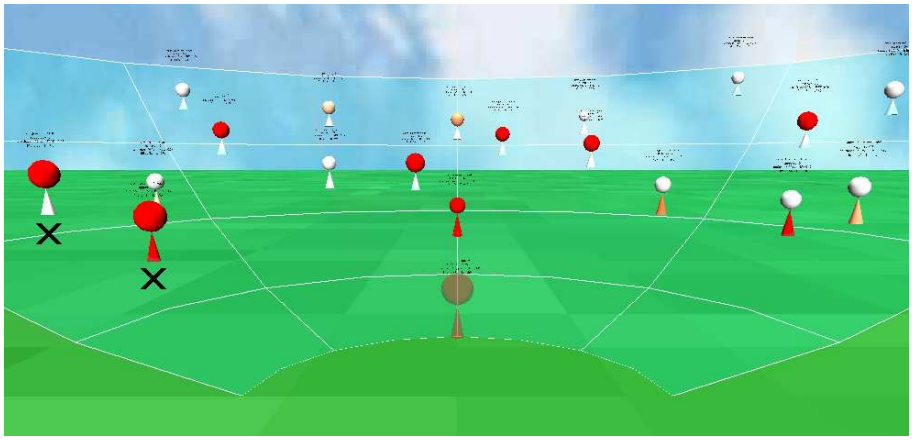


FIG. 5.15.d

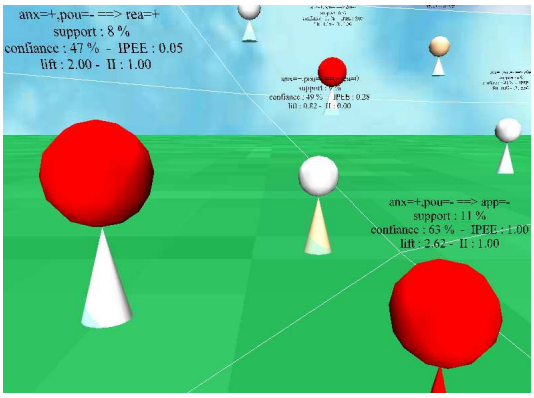


FIG. 5.15.e

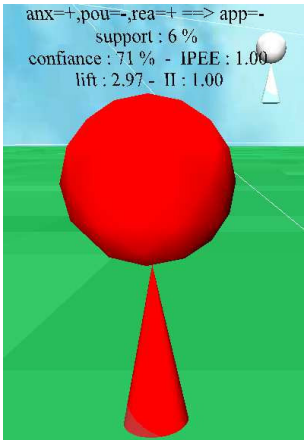


FIG. 5.15.f

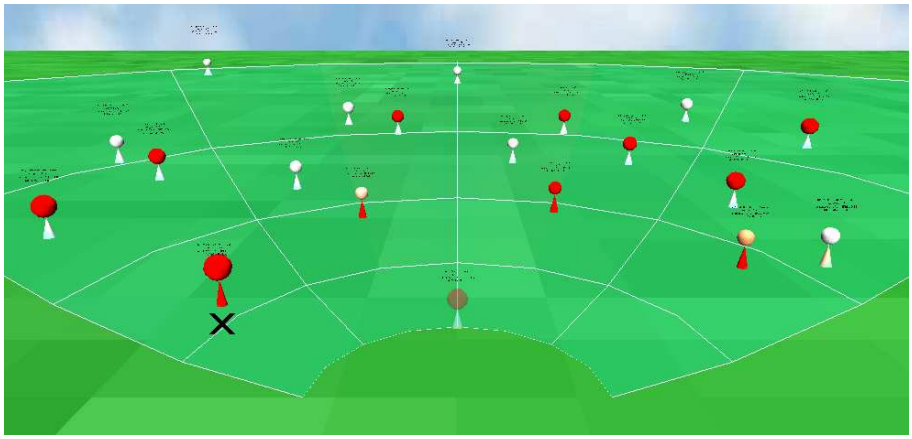


FIG. 5.15.g

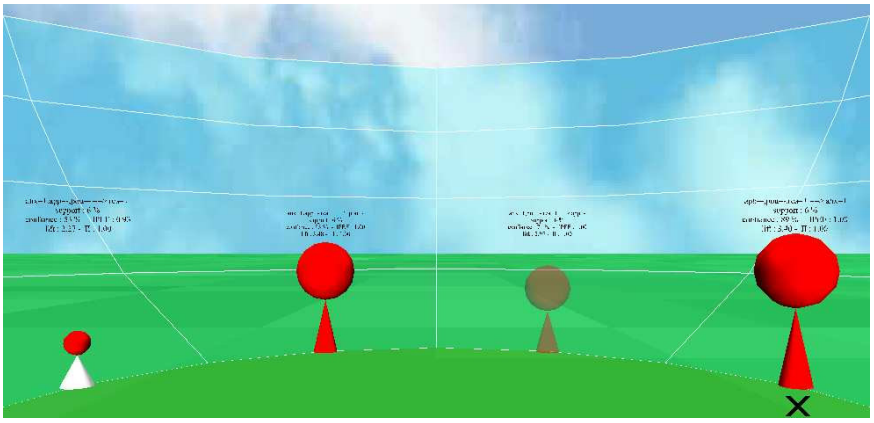


FIG. 5.15.h

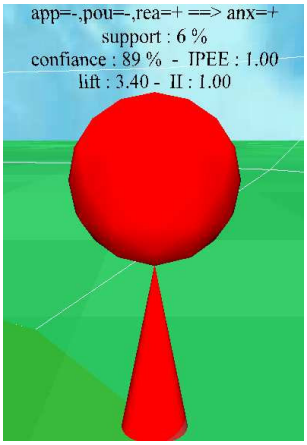


FIG. 5.15.i

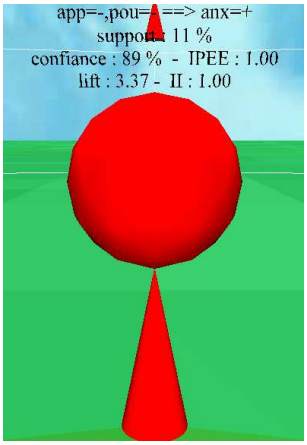


FIG. 5.15.j

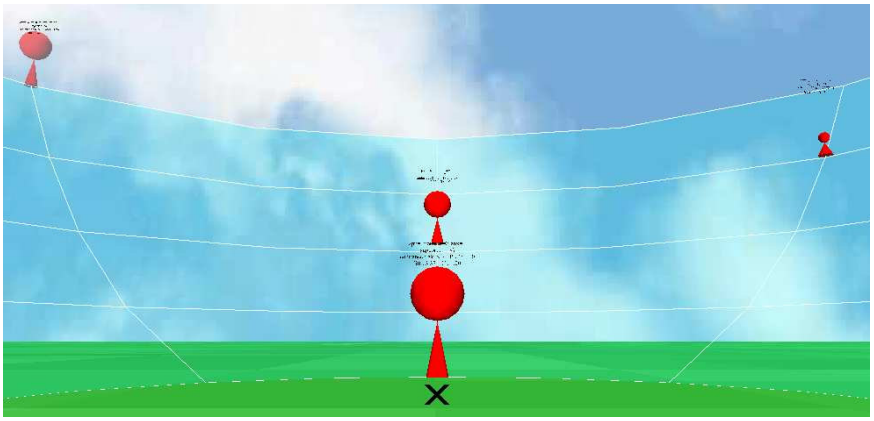


FIG. 5.15.k

FIG. 5.15 – Paysages de l'exemple 1

Nous commençons par nous intéresser aux traits comportementaux impliqués par la motivation de protection (personnes qui recherchent la sécurité). Pour cela, nous visualisons le sous-ensemble des règles qui possèdent l'itemset $(pou = -)$ en prémisse (figure 5.15.a). Dans le paysage, une règle est clairement mise en évidence avec un objet volumineux et rouge placé à l'avant de l'arène. Il s'agit de $(pou = -) \rightarrow (anx = +)$, qui possède en effet de très bonnes valeurs d'indices de règle (figure 5.15.b). En visitant le paysage, nous constatons qu'il s'agit de la meilleure règle du sous-ensemble. En particulier, les objets constitués d'un cône rouge et d'une sphère blanche désignent des règles dont l'écart à l'équilibre est significatif et orienté en faveur des exemples (IPEE élevé) alors que l'écart à l'indépendance est significatif mais orienté en faveur des contre-exemples (intensité d'implication faible). Similairement, les objets constitués d'un cône blanc et d'une sphère rouge désignent des règles dont l'écart à l'indépendance est significatif et orienté en faveur des exemples (intensité d'implication élevée) alors que l'écart à l'équilibre est significatif mais orienté en faveur des contre-exemples (IPEE faible). Les luminosités intermédiaires entre blanc et rouge représentent quant à elles des écarts non significatifs (cas peu fréquent). Afin de savoir s'il existe d'autres règles qui prédisent l'anxiété plus efficacement, nous appliquons sur $(pou = -) \rightarrow (anx = +)$ la relation *Conclusion commune*.

Dans le nouveau paysage (figure 5.15.c), la règle de transition $(pou = -) \rightarrow (anx = +)$ est située en bas à gauche de l'arène, représentée par un objet semi-transparent. La seule autre règle qui présente une qualité comparable (objet rouge derrière la règle de transition) possède une confiance de 60%, ce qui est relativement faible. Tous les autres objets sont constitués d'un cône blanc (IPEE faible). Ce paysage nous permet donc de constater que la motivation de protection est le meilleur caractère qui implique l'anxiété. Afin d'en savoir plus sur cette population anxieuse qui recherche la sécurité, nous appliquons sur $(pou = -) \rightarrow (anx = +)$ la relation *Chaînage avant*. Le nouveau paysage présente toutes les règles qui possèdent $(anx = +, pou = -)$ en prémisse (figure 5.15.d). La règle de transition mise à part, les écarts à l'équilibre sont mauvais ou faibles, mais nous pouvons tenter de maximiser les écarts à l'indépendance. Pour cela, la vue du dessus est pratique. Deux règles sortent du lot : $(anx = +, pou = -) \rightarrow (app = -)$ et $(anx = +, pou = -) \rightarrow (rea = +)$, qui possèdent les valeurs de lift les plus élevées (2.62 et 2.00 respectivement, voir figure 5.15.e). La règle $(anx = +, pou = -) \rightarrow (rea = +)$ concerne la population anxieuse, motivée par la protection, et motivée par la réalisation (personnes motivées par la création et l'effort plus que par les objectifs comme l'argent ou la reconnaissance). Afin de découvrir les caractères impliqués par cette population, tout en essayant d'améliorer la qualité des règles, nous appliquons la relation *Chaînage avant* sur $(anx = +, pou = -) \rightarrow (rea = +)$.

Le sous-ensemble suivant (figure 5.15.g) contient une règle très bonne et largement meilleure que les autres : $(anx = +, pou = -, rea = +) \rightarrow (app = -)$ (figure 5.15.f). Il est à noter que l'item sur lequel la règle conclut (motivation d'indépendance) intervenait déjà dans l'une des deux meilleures règles du paysage précédent. Après avoir spécialisé les règles en ajoutant des items, nous allons maintenant nous concentrer sur ces quatre items et passer à une phase de généralisation. Tout d'abord, nous appliquons la relation *Items communs* afin de vérifier si une combinaison différente des items ne serait pas plus pertinente. En comparant à la règle de transition, nous pouvons voir sur le nouveau paysage

que deux combinaisons sont largement meilleures (figure 5.15.h). La meilleure d'entre elles est la règle $(app = -, pou = -, rea = +) \rightarrow (anx = +)$ avec une confiance de 89% et un lift de 3.40 (validés par des indices probabilistes au maximum), ce qui est tout à fait satisfaisant (figure 5.15.i). Afin de vérifier que tous les items en prémisse sont bien utiles pour prédire l'anxiété, nous appliquons la relation *Généralisation*. Dans le nouveau paysage (figure 5.15.k), la meilleure règle est quasiment aussi bonne que la précédente : il s'agit de $(app = -, pou = -) \rightarrow (anx = +)$ avec une confiance de 89% et un lift de 3.37 (figure 5.15.j). L'item $rea = +$ était donc superflu. Nous sommes parvenu à simplifier la règle (et donc à augmenter le support, qui est passé de 6% à 11%) tout en préservant les valeurs des indices de règle pour la prédiction de l'anxiété. En appliquant la relation *Généralisation* une nouvelle fois, nous constatons que tous les items de la règle sont utiles, puisque les règles $(app = -) \rightarrow (anx = +)$ et $(pou = -) \rightarrow (anx = +)$ sont nettement plus mauvaises (ces règles ont déjà été visualisées dans le paysage de la figure 5.15.c). Au final, nous retenons donc la règle $(app = -, pou = -) \rightarrow (anx = +)$, qui signifie que les personnes qui recherchent l'indépendance et la sécurité sont généralement anxieuses. L'exploration réalisée avec *ARVis* a montré non seulement que cette règle est de très bonne qualité, mais aussi qu'elle est localement dominante (la meilleure dans la "région" de règles explorée).

5.5.2 Exemple 2

Dans ce nouvel exemple, nous cherchons à prédire l'affirmation de soi. Pour cela, nous visualisons le sous-ensemble des règles qui concluent sur $aff = +$ (figure 5.16.a). Dans le paysage, une règle saute aux yeux : il s'agit de $(com = -) \rightarrow (aff = +)$, avec une confiance moyenne de 70% et un bon lift de 2.85 (règle indiquée par une croix sur la figure 5.16.a). Afin de visualiser des règles plus spécifiques qui prédisent l'affirmation de soi, nous appliquons sur $(com = -) \rightarrow (aff = +)$ la relation de voisinage *Spécialisation concordante*. Le nouveau paysage contient beaucoup de règles de bonne qualité (figure 5.16.b). En particulier, tous les écarts à l'indépendance et à l'équilibre sont significativement élevés (tous les objets sont rouges). Pour faciliter la visualisation des meilleures règles, nous filtrons le sous-ensemble avec un seuil minimal de lift égal à 3. Le paysage obtenu met en évidence la règle $(com = -, rig = -) \rightarrow (aff = +)$ (figure 5.16.c), dont le support est plus élevé que les autres (7% contre environ 3% pour les autres au fond de l'arène). En plus d'un lift supérieur à 3, cette règle possède une confiance convenable de 76%.

En appliquant à nouveau *Spécialisation concordante*, nous obtenons des règles plus spécifiques qui prédisent l'affirmation de soi (figure 5.16.d). Toutefois, ces règles possèdent toutes un très faible support (de l'ordre de 0.1%, c'est-à-dire environ quatre profils), ce qui nous incite à mettre un terme à la spécialisation des règles. Il n'est pas nécessaire de généraliser puisque les règles $(com = -) \rightarrow (aff = +)$ et $(rig = -) \rightarrow (aff = +)$ ont déjà été visualisées dans le premier paysage (figure 5.16.a). Au final, nous validons la règle $(com = -, rig = -) \rightarrow (aff = +)$, qui signifie que les personnes qui recherchent la conciliation et s'adaptent bien à l'imprévu ont tendance à s'affirmer. D'après la méthodologie d'évaluation de *PerformanSe SA*, et en l'absence d'informations

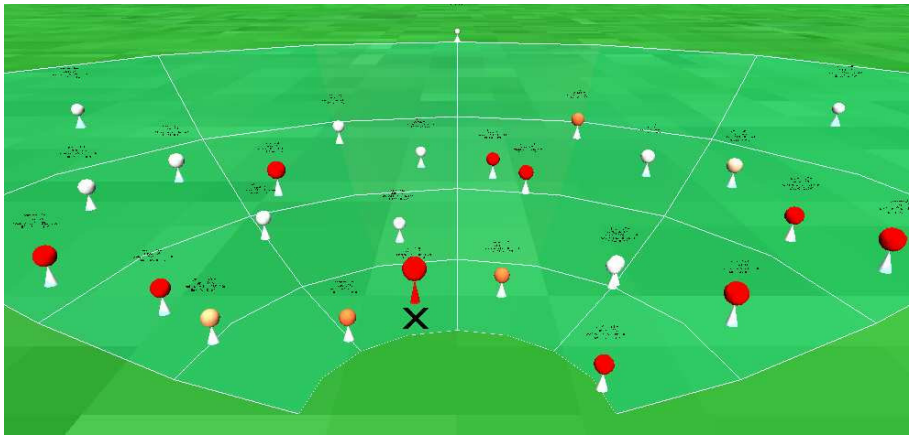


FIG. 5.16.a

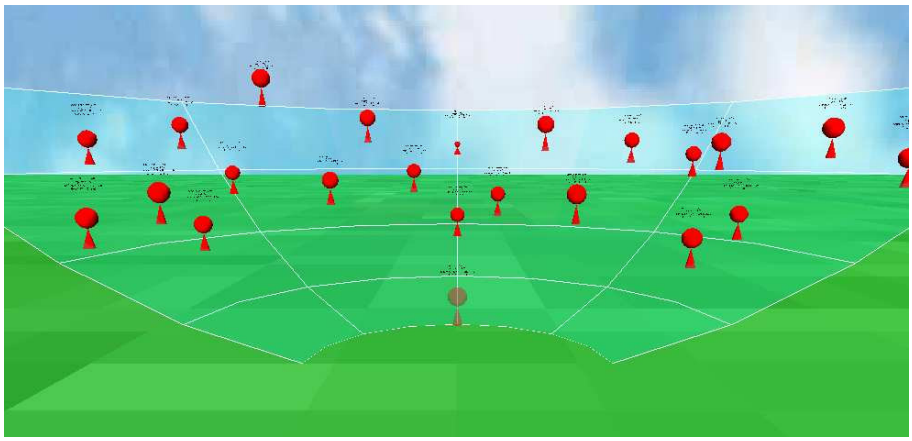


FIG. 5.16.b

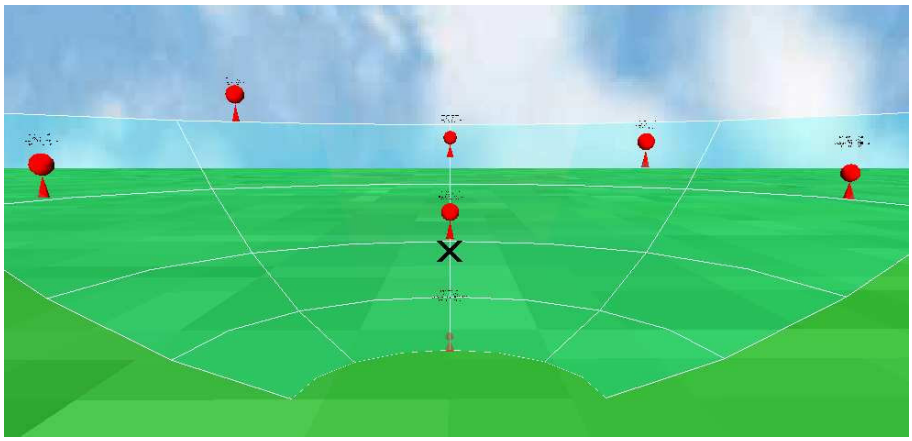


FIG. 5.16.c

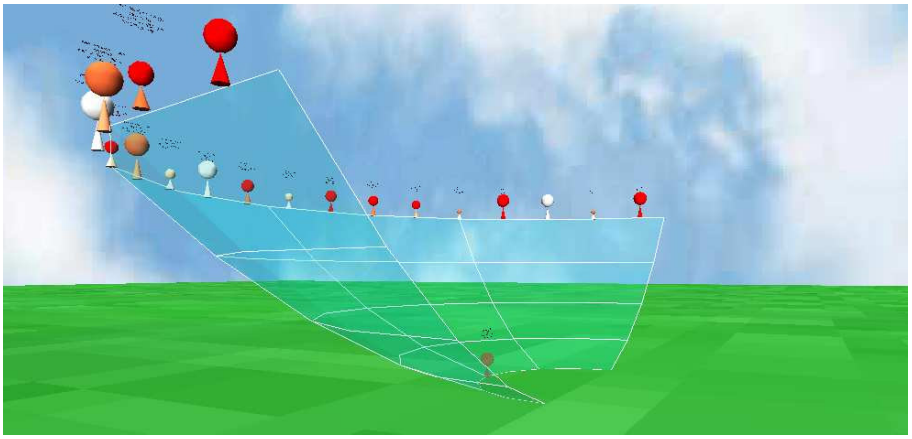


FIG. 5.16.d

FIG. 5.16 – Paysages de l'exemple 2

supplémentaires, ces personnes peuvent être décrites a priori comme travaillant efficacement en environnement incertain, sachant prendre du recul, mais pouvant prendre des décisions risquées.

5.5.3 Exemple 3

Nous voulons ici vérifier la règle selon laquelle les personnes qui recherchent l'indépendance sont rigoureuses : $(app = -) \rightarrow (rig = +)$. Pour cela, nous visualisons le sous-ensemble des règles qui concluent sur $rig = +$. Le paysage montre de nombreuses règles de qualité (figure 5.17.a). A l'aide du menu proposant la liste des règles, nous affichons directement le point de vue rapproché sur la règle qui nous intéresse (figure 5.17.b). Il s'agit d'une règle de très faible qualité, avec de mauvais écarts à l'équilibre et à l'indépendance validés par les indices probabilistes. La liaison entre les traits comportementaux *app* et *rig* réside en fait davantage dans la règle contraire $(app = -) \rightarrow (\overline{rig = +})$.

Afin de tenter d'améliorer la prédiction de $rig = +$ par $app = -$, nous appliquons la relation de voisinage *Spécialisation concordante*. Dans le nouveau paysage (figure 5.17.d), la meilleure règle est $(app = -, pou = +) \rightarrow (rig = +)$, qui possède en particulier une confiance élevée confortée par une valeur de IPEE au maximum (figure 5.17.e). Cette règle est intéressante, mais avant de la valider il faut examiner la règle plus générale $(pou = +) \rightarrow (rig = +)$ afin de vérifier que $pou = +$ n'est pas à lui seul un bon prédicteur de $rig = +$. En revenant au paysage précédent, nous pouvons vérifier que cette règle est de qualité moyenne (figure 5.17.c). La bonne prédiction de $rig = +$ dépend donc pleinement des deux caractères $app = -$ et $pou = +$. Au final, nous retenons la règle $(app = -, pou = +) \rightarrow (rig = +)$, qui signifie que les personnes qui recherchent l'indépendance et sont motivées par le pouvoir sont généralement rigoureuses. Par rapport aux autre règles du jeu de données, il s'agit d'une règle très spécifique qui explique une niche dans la population.

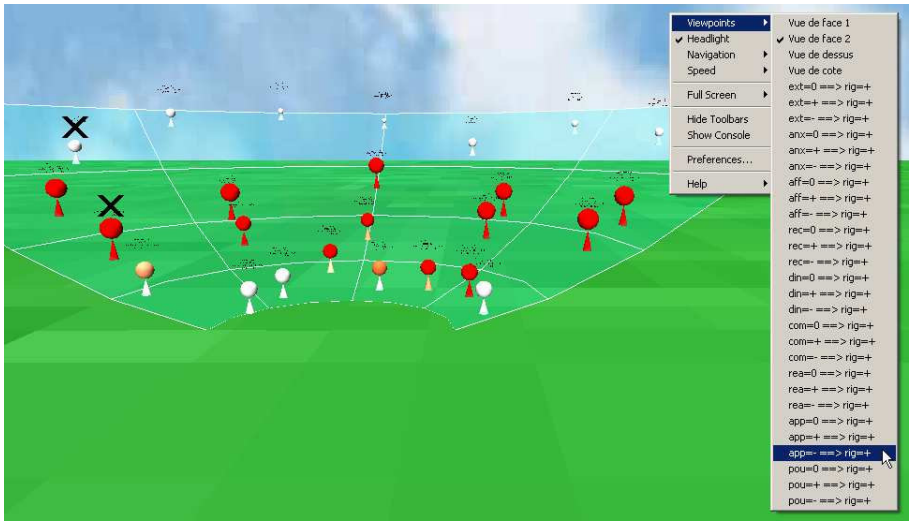


FIG. 5.17.a

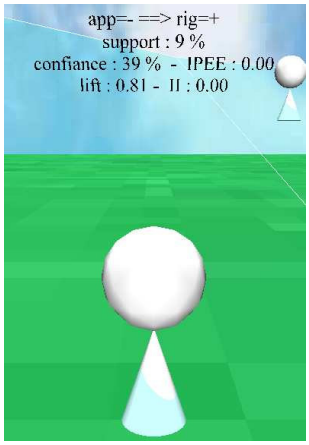


FIG. 5.17.b

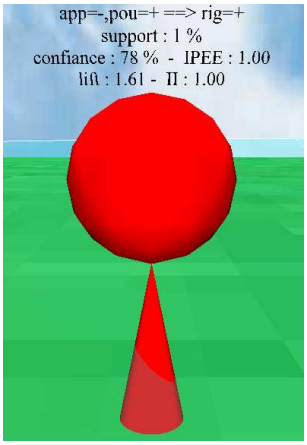


FIG. 5.17.c

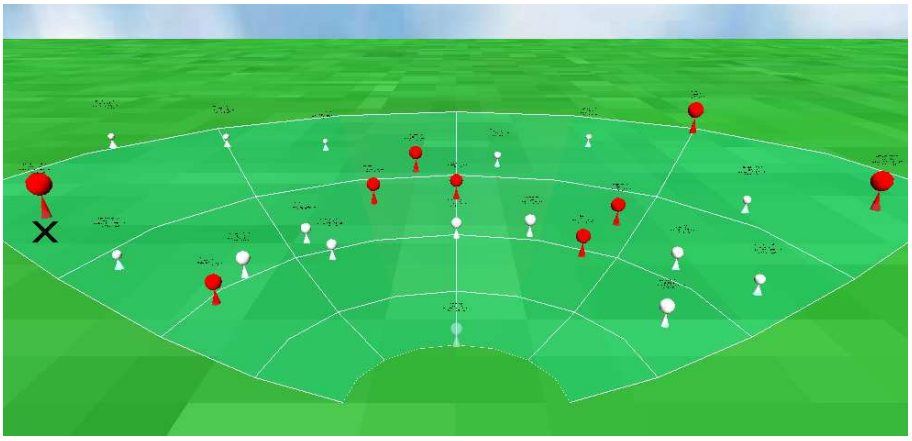


FIG. 5.17.d

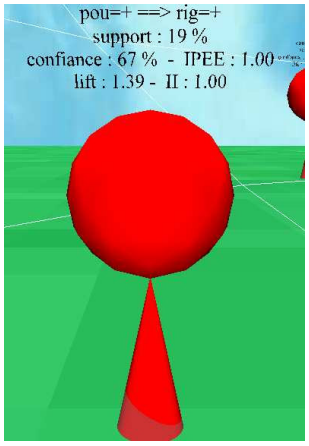


FIG. 5.17.e

FIG. 5.17 – Paysages de l'exemple 3

5.6 Conclusion

Dans ce chapitre, nous avons présenté l'outil de visualisation *ARVis* (*Association Rule Visualization*), une implémentation de la méthodologie *RF*. *ARVis* permet d'explorer de grands volumes de règles et d'identifier les connaissances pertinentes. L'outil repose sur une représentation 3D intuitive qui supporte de grands ensembles de règles décrits par plusieurs indices. Fondée sur la métaphore du paysage d'information, cette représentation inédite en visualisation de règles met en évidence les indices et facilite la reconnaissance des règles de bonne qualité. Pour ce qui concerne les interactions, plusieurs relations de voisinage intelligibles réalisant des spécialisations ou généralisations permettent à l'utilisateur de guider sa navigation parmi les règles. Les tests d'*ARVis* sur données réelles montrent que l'outil aide à découvrir des règles intéressantes et de bonne qualité, et en particulier des règles localement dominantes (les meilleures dans la "région" de règles explorée).

Comme l'a montré le chapitre 4, les méthodes et outils de visualisation de règles sont fondés sur les items et considèrent les indices de règle comme des informations annexes. La philosophie d'*ARVis* est toute autre :

- la représentation est principalement fondée sur les indices de règle (les items apparaissent uniquement dans des étiquettes textuelles) ;
- par le biais des relations de voisinage, l'interaction est principalement fondée sur les items et permet de jouer sur la syntaxe des règles.

Extraction locale interactive des règles avec *ARVis* 6

Sommaire

Dans ce dernier chapitre, nous présentons le module d'extraction de règles d'*ARVis*. Il s'agit d'un module interactif qui produit à la demande les ensembles de règles (et les valeurs d'indices) que l'utilisateur souhaite visualiser. Aucune règle n'est donc produite à l'avance. Cette *approche locale* qui suit la navigation de l'utilisateur permet de s'affranchir des limites de l'extraction exhaustive, c'est-à-dire principalement des temps d'exécution qui peuvent s'avérer rédhibitoires sur des données denses, quand l'explosion combinatoire ne rend pas l'extraction impossible (voir section 3.5 page 81).

L'extraction locale des règles est réalisée par des algorithmes à contraintes, conçus spécifiquement pour extraire à la demande les ensembles de règles que les relations de voisinage engendrent. Les algorithmes à contraintes d'*ARVis* ont été écrits dans une optique tout à fait différente des algorithmes à contraintes présentés au chapitre 3. Alors que ces derniers cherchent à exploiter des classes de contraintes les plus générales possibles, les algorithmes d'*ARVis* utilisent uniquement les contraintes particulières induites par les relations de voisinage, et sont spécialement optimisés pour ces contraintes.

Nous décrivons deux approches pour l'extraction locale des règles :
- l'extraction *sans mémoire*, approche basique qui ne sauvegarde pas les itemsets générés (implémentée dans *ARVis 1.1*) ;
- l'extraction *avec mémoire*, approche améliorée qui sauvegarde les itemsets et leurs cardinaux pour pouvoir les réutiliser (implémentée dans *ARVis*

1.2).

Dans la section suivante, nous présentons les contraintes utilisées dans *ARVis*. Les sections 6.2 et 6.3 sont ensuite consacrées aux algorithmes d'extraction sans mémoire et avec mémoire d'*ARVis 1.1* et *1.2*. Enfin, dans la section 6.4, nous étudions les temps de réponses d'*ARVis 1.2*. Tout au long du chapitre, les relations de voisinage considérées sont celles d'*ARVis 1.2* puisqu'elles incluent celles d'*ARVis 1.1*.

6.1 Contraintes dans *ARVis*

Les algorithmes à contraintes utilisent les contraintes pour réduire l'espace de recherche (voir chapitre 3). Dans *ARVis*, les relations de voisinage engendrent deux types de contraintes sur les règles :

- des contraintes syntaxiques, qui indiquent quels items peuvent ou doivent apparaître en prémisse et conclusion ;
- des contraintes de qualité, qui spécifient un seuil minimal et un seuil maximal pour chaque indice de règle.

Les contraintes syntaxiques d'*ARVis* sont des contraintes puissantes qui limitent drastiquement l'espace de recherche. En effet, si I est l'ensemble des items décrivant les données, alors les contraintes syntaxiques sont vérifiées par au plus $|I|$ règles. Ainsi, quelle que soit la relation de voisinage choisie par l'utilisateur, les algorithmes d'extraction de règles d'*ARVis* ont une complexité polynomiale en $O(|I|)$ [1].

6.2 Extraction locale sans mémoire (*ARVis 1.1*)

L'algorithme pour l'extraction locale sans mémoire est l'algorithme 6.1. Il est suffisamment général pour être commun à toutes les relations de voisinage (en particulier, même s'il n'a été implémenté que pour les deux relations de voisinage d'*ARVis 1.1*, il est également utilisable pour toutes les relations d'*ARVis 1.2*). Cet algorithme est organisé en quatre étapes.

- L'étape 1 utilise les contraintes syntaxiques. Comme nous l'avons vu au chapitre 3, les contraintes dites syntaxiques disposent toutes d'une fonction génératrice, c'est-à-dire qu'il est possible d'énumérer toutes les règles qui vérifient une contrainte syntaxique uniquement à partir de la liste des items, sans avoir à consulter les individus dans les données (voir section 3.4.2 page 79). C'est ce qu'effectue l'étape 1, qui ne requiert ainsi aucune lecture dans la base de données. C'est aussi la seule étape de l'algorithme qui dépende de la relation de voisinage Π choisie par l'utilisateur.
- L'étape 2 consiste à déterminer le cardinal de chaque itemset (nombre d'individus qui vérifient l'itemset dans les données) apparaissant dans les règles énumérées à l'étape 1. Pour une règle $a \rightarrow b$, il faut considérer

[1] Sauf pour la relation *Spécialisation d'exception* pour laquelle le nombre de règles est borné par $m|I|$ et la complexité est en $O(m|I|)$, où $m < |I|$ est le nombre maximum de modalités pour une variable.

Entrées : – *règle*, la règle de transition,
– Π, la relation de voisinage,
– *seuils*, les seuils min/maximaux sur les indices de règle,
– $\mathcal{BD}$, base de données

Sorties : – $ensemble_{règles+indices}$, ensemble de règles avec les valeurs des indices

```
1  ensemble_règles = ∅ ;          //ensemble de règles sans indice
2  cardinaux = ∅ ;                //cardinaux des itemsets

3  //ETAPE 1 : Construire les règles candidates avec les
   contraintes syntaxiques
4  ensemble_règles = constructionSyntaxique(règle, Π) ;

5  //ETAPE 2 : Compter dans la base de données les occurrences
   des itemsets des règles candidates
6  cardinaux = compterItemsets(ensemble_règles, BD) ;

7  //ETAPE 3 : Calculer les indices de règle
8  ensemble_règles+indices = calculerIndic(ensemble_règles, cardinaux) ;

9  //ETAPE 4 : Eliminer les règles candidates qui ne respectent
   pas les contraintes de qualité
10 ensemble_règles+indices = filtrer(ensemble_règles+indices, seuils) ;

11 retourne ensemble_règles+indices ;
   (utilisable pour toutes les relations de voisinage)
```

Algorithme 6.1: Algorithme pour l'extraction locale sans mémoire

trois itemsets : la prémisse (de cardinal n_a), la conclusion (n_b), et l'itemset global [prémisse $\cup$ conclusion] (n_{ab}). Les cardinaux sont déterminés en comptant les occurrences des itemsets dans la base de données, ce qui fait de cette étape la plus coûteuse en temps de l'algorithme 6.1.

- L'étape 3 consiste à calculer les valeurs des indices de règle à partir des cardinaux n_a, n_b, et n_{ab} déterminés à l'étape 2 (le cardinal n du jeu de données est aussi nécessaire).
- A l'étape 4 sont exploitées les contraintes de qualité : les règles sont filtrées sur les valeurs des indices en fonction des seuils.

6.3 Extraction locale avec mémoire (*ARVis 1.2*)

L'extraction locale avec mémoire reprend les quatre étapes de l'extraction locale sans mémoire (algorithme 6.1), mais avec les deux améliorations suivantes.

- Pour améliorer les temps de réponse, un dispositif de sauvegarde progressive des itemsets est implémenté pour éviter d'avoir à compter plusieurs fois le même itemset dans la base de données.

- Les contraintes de qualité sont exploitées (dans la mesure du possible) dès l'étape 1 pour réduire encore davantage l'espace de recherche et donc diminuer le nombre d'itemsets à compter dans la base de données. Ceci est rendu possible grâce à la sauvegarde progressive des itemsets : comme les cardinaux sauvés sont accessibles sans avoir à lire les données, ils peuvent être exploités directement à l'étape 1 de construction syntaxique des règles, pour anticiper que certaines règles candidates ne respectent pas les contraintes de qualité.

6.3.1 Sauvegarde progressive des itemsets

Entrées : – *itemset*,
– $\mathcal{BD}$, base de données
Sorties : – *cardinal*, cardinal de l'itemset

Procédure `récupérerCardinal()`

```
1 si (itemset déjà sauvegardé) alors
2 |  retourne itemset.cardinal ;
3 sinon
4 |  cardinal = compterDansLaBase(itemset, BD) ;
5 |                                  //une passe sur les données
6 |  sauvegarder(itemset, cardinal, BD) ;
7 |                                  //sauvegarde progressive
8 |  retourne cardinal ;
```

Algorithme 6.2: Procédure `récupérerCardinal()`

Le dispositif de sauvegarde progressive des itemsets intervient à l'étape 2 : chaque fois qu'un itemset a été compté, son cardinal est sauvegardé en base de données pour éviter d'avoir à le déterminer une nouvelle fois durant la suite de l'exploration. De cette façon, plus *ARVis* est utilisé sur un jeu de données, plus les cardinaux des itemsets sont sauvegardés, et plus l'algorithme d'extraction locale a de chances de s'exécuter rapidement. Plus précisément, les itemsets et leurs cardinaux sont sauvegardés dans la base de données dans des tables spécifiques. Il existe une table par longueur d'itemset, de telle façon que tous les itemsets de même longueur (possédant le même nombre d'items) sont stockés dans la même table. Chaque table est munie d'un index B-arbre sur les itemsets.

C'est par la procédure `récupérerCardinal()` (algorithme 6.2) qu'est implémenté le dispositif de sauvegarde progressive des itemsets. Cette procédure est appelée dans les algorithmes d'extraction avec mémoire pour obtenir les cardinaux des itemsets. Avant de compter les occurrences d'un itemset dans la base de données, elle vérifie si son cardinal n'a pas déjà été déterminé.

6.3.2 Optimisation des contraintes de qualité

L'algorithme sans mémoire 6.1 est suffisamment général pour concerner n'importe quelle relation de voisinage. En revanche, l'extraction avec mémoire dépend fortement de la relation de voisinage considérée. Nous allons donc détailler plusieurs algorithmes pour l'extraction locale avec mémoire. Plus précisément, nous distinguons :

1. les algorithmes qui produisent des règles en utilisant uniquement les items de la règle de transition (pour les relations de voisinage *Généralisation*, *Généralisation de la prémisse*, et *Items communs*) ;
2. les algorithmes qui produisent des règles en insérant un nouvel item dans la règle de transition (pour les relations de voisinage *Spécialisation concordante*, *Spécialisation d'exception*, *Chaînage avant*, *Prémisse commune*, et *Conclusion commune*).

Les algorithmes de la première catégorie sont calqués sur l'algorithme sans mémoire 6.1 : ils utilisent la sauvegarde progressive des itemsets (à l'étape 2) mais ne présentent aucune optimisation pour les contraintes de qualité. Ces algorithmes produisent en effet peu de règles et s'exécutent très rapidement ; ce n'est pas sur eux que doit se porter l'effort d'optimisation. Par exemple, si l'utilisateur n'étudie que des règles comprenant moins de cinq items (ce qui semble raisonnable dans la pratique), alors les algorithmes de la première catégorie ne produisent jamais plus de cinq règles.

Les algorithmes de la seconde catégorie, en revanche, peuvent produire beaucoup plus de règles. Ils sont optimisés à l'étape 1 de la manière suivante : les items à insérer dans la règle de transition sont sélectionnés en fonction des contraintes de qualité. Ceci permet d'anticiper que certains items, une fois insérés dans la règle de transition, mènent à des règles candidates qui ne respectent pas les contraintes de qualité.

Exemple. Avec la relation de voisinage *Conclusion commune*, il est inutile de générer à l'étape 1 une règle candidate $a \rightarrow b$ si le cardinal n_a a déjà été déterminé et qu'il ne vérifie pas $n_a \geq \frac{n.minSupport}{maxConf}$. En effet, les contraintes de qualité

imposent

$$minSupport \leq \frac{n_{ab}}{n} \quad \text{et} \quad \frac{n_{ab}}{n_a} \leq maxConf,$$

dont on déduit

$$n_a \geq \frac{n_{ab}}{maxConf} \geq \frac{n.minSupport}{maxConf} \ . \qquad \square$$

L'avantage d'utiliser le niveau 1 (itemsets de longueur 1, c'est-à-dire les items) par rapport aux autres niveaux pour effectuer ce filtrage sur les règles candidates est que la totalité du niveau 1 est généralement connue du dispositif de sauvegarde progressive, puisque :

- de par la combinatoire, le niveau 1 est plus restreint que les niveaux qui lui sont supérieurs ;
- de par les relations de voisinage choisies dans *ARVis*, le niveau 1 est fréquemment sollicité par les algorithmes d'extraction.

Dans la suite, nous donnons une description plus détaillée de ces algorithmes avec mémoire qui sont optimisés à l'étape 1.

Spécialisation concordante

L'algorithme 6.3 permet d'extraire les ensembles de règles délimités par la relation de voisinage *Spécialisation concordante* (Π_1 selon les notations du chapitre 5). Les contraintes de qualité sont exploitées dès l'étape 1 à la ligne 9. Cet algorithme est également utilisé pour la relation *Spécialisation d'exception* (Π_2). Par exemple, avec la variable *couleur_des_yeux* et les modalités *marron*, *vert*, *bleu*, on a :

$$\Pi_2\Big((couleur\_des\_cheveux = blond) \rightarrow (couleur\_des\_yeux = bleu)\Big)$$

$$=$$

$$\Pi_1\Big((couleur\_des\_cheveux = blond) \rightarrow (couleur\_des\_yeux = vert)\Big)$$

$$\cup$$

$$\Pi_1\Big((couleur\_des\_cheveux = blond) \rightarrow (couleur\_des\_yeux = marron)\Big)$$

Chaînage avant

L'algorithme 6.4 permet d'extraire les ensembles de règles délimités par la relation de voisinage *Chaînage avant* (Π_3). Les contraintes de qualité sont exploitées dès l'étape 1 à la ligne 6. Cet algorithme est également utilisé pour la relation *Prémisse commune* (Π_6). Par exemple, avec trois items A, B, et C, on a :

$$\Pi_6\Big((A, B) \rightarrow (C)\Big) = \Pi_3\Big((A) \rightarrow (B)\Big) = \Pi_3\Big((B) \rightarrow (A)\Big)$$

Entrées : – *prémisse*, prémisse de la règle de transition,
– *conclusion*, conclusion de la règle de transition,
– $nbreIndividus$, nombre total d'individus dans les données,
– $minSupport$, seuil de support minimal,
– $maxSupport$, seuil de support maximal,
– $minConf$, seuil de confiance minimale,
– $maxConf$, seuil de confiance maximale,
– $minLift$, seuil de lift minimal,
– $maxLift$, seuil de lift maximal,
– $minIPEE$, seuil d'IPEE minimal,
– $maxIPEE$, seuil d'IPEE maximal,
– $minII$, seuil d'intensité d'implication minimale,
– $maxII$, seuil d'intensité d'implication maximale,
– $\mathcal{BD}$, base de données

Sorties : – *ensemble*, ensemble de règles avec les valeurs des indices

$ensemble = \emptyset$;
item ; `//item à insérer dans la règle de transition`
cardinal ; `//cardinal de` *item*
règle ; `//règle créée en insérant` *item*
`//ETAPE 1`
pour chaque $item \in \mathcal{BD}$ **faire**
  si *(item ∉ prémisse* **et** *item ∉ conclusion)* **alors**
    $cardinal =$ `récupérerCardinal(`$item, \mathcal{BD}$`)` ;
    si $(cardinal \geq nbreIndividus \times minSupport \div maxConf$
    et $cardinal \geq$
    $nbreIndividus^2 \times minSupport \div conclusion.cardinal \div maxLift$
    et $cardinal \geq$
    $conclusion.cardinal \times minLift \times minSupport \div maxConf)$
    alors
      `//ETAPE 2`
      `récupérerCardinal(`$prémisse \cup item, \mathcal{BD}$`)` ;
      `récupérerCardinal(`$prémisse \cup conclusion \cup item, \mathcal{BD}$`)` ;
      `//ETAPE 3`
      $règle = (prémisse \cup item \rightarrow conclusion)$;
      $support =$ `calculerSupport(`$règle$`)` ;
      $confiance =$ `calculerConfiance(`$règle$`)` ;
      $lift =$ `calculerLift(`$règle$`)` ;
      $IPEE =$ `calculerIPEE(`$règle$`)` ;
      $II =$ `calculerII(`$règle$`)` ;
      `//ETAPE 4`
      si $(minSupport \leq support \leq maxSupport$
      et $minConf \leq confiance \leq maxConf$
      et $minLift \leq lift \leq maxLift$
      et $minIPEE \leq IPEE \leq maxIPEE$
      et $minII \leq II \leq maxII)$ **alors**
        $ensemble = ensemble \cup \left\{(règle, support, confiance, lift, IPEE, II)\right\}$;

retourne *ensemble* ;

Algorithme 6.3: Algorithme d'extraction locale avec mémoire pour la relation *Spécialisation concordante*

```
1  ensemble = ∅ ;
2  //ETAPE 1
3  pour chaque item ∈ BD faire
4      si (item ∉ prémisse et item ∉ conclusion) alors
5          cardinal = récupérerCardinal(item, BD) ;
6          si (cardinal ≥ nbreIndividus × minSupport
           et cardinal ≥ minConf × prémisse.cardinal
           et cardinal ≥
           nbreIndividus² × minSupport ÷ prémisse.cardinal ÷ maxLift
           et cardinal ≥ nbreIndividus × minConf ÷ maxLift
           et cardinal ≤
           nbreIndividus² × maxSupport ÷ prémisse.cardinal ÷ minLift
           et cardinal ≤ nbreIndividus × maxConf ÷ minLift) alors
7              //ETAPE 2
8              récupérerCardinal(prémisse ∪ conclusion ∪ item, BD) ;
9              //ETAPE 3
10             règle = (prémisse ∪ conclusion → item) ;
11             support = calculerSupport(règle) ;
12             confiance = calculerConfiance(règle) ;
13             lift = calculerLift(règle) ;
14             IPEE = calculerIPEE(règle) ;
15             II = calculerII(règle) ;
16             //ETAPE 4
17             si (minSupport ≤  support  ≤ maxSupport
               et  minConf ≤ confiance ≤ maxConf
               et  minLift ≤  lift  ≤ maxLift
               et  minIPEE ≤  IPEE  ≤ maxIPEE
               et  minII ≤  II  ≤ maxII) alors
18                 ensemble = ensemble ∪
                   {(règle, support, confiance, lift, IPEE, II)} ;
19 retourne ensemble ;
```

Algorithme 6.4: Algorithme d'extraction locale avec mémoire pour la relation *Chaînage avant*

```
1  ensemble = ∅ ;
2  //ETAPE 1
3  pour chaque item ∈ BD faire
4      si (item ∉ conclusion) alors
5          cardinal = récupérerCardinal(item, BD) ;
6          si (cardinal ≥ nbreIndividus × minSupport ÷ maxConf
           et cardinal ≥
           nbreIndividus² × minSupport ÷ conclusion.cardinal ÷ maxLift
           et cardinal ≤ nbreIndividus × maxSupport ÷ minConf
           et cardinal ≤
           nbreIndividus² × maxSupport ÷ conclusion.cardinal ÷ minLift)
           alors
7              //ETAPE 2
8              récupérerCardinal(conclusion ∪ item, BD) ;
9              //ETAPE 3
10             règle = (item → conclusion) ;
11             support = calculerSupport(règle) ;
12             confiance = calculerConfiance(règle) ;
13             lift = calculerLift(règle) ;
14             IPEE = calculerIPEE(règle) ;
15             II = calculerII(règle) ;
16             //ETAPE 4
17             si (minSupport ≤ support ≤ maxSupport
               et minConf ≤ confiance ≤ maxConf
               et minLift ≤ lift ≤ maxLift
               et minIPEE ≤ IPEE ≤ maxIPEE
               et minII ≤ II ≤ maxII) alors
18                 ensemble = ensemble ∪
                   {(règle, support, confiance, lift, IPEE, II)} ;
19 retourne ensemble ;
```

Algorithme 6.5: Algorithme d'extraction locale avec mémoire pour la relation *Conclusion commune*

	Nombre d'items	**Nombre d'individus**	**Nombre moyen d'items vrais par individu**
MUSHROOMS	119	8416	23
T10.I4.D100k	100	100000	10
T20.I6.D100k	40	100000	20

TAB. 6.1 – Caractéristiques des données

Conclusion commune

L'algorithme 6.5 permet d'extraire les ensembles de règles délimités par la relation de voisinage *Conclusion commune*. Les contraintes de qualité sont exploitées dès l'étape 1 à la ligne 6.

6.4 Temps de réponse dans *ARVis 1.2*

La figure 6.1 montre les temps de réponse obtenus sur trois jeux de données (présentés tableau 6.1) en exécutant un scénario d'exploration avec *ARVis 1.2*, c'est-à-dire une suite d'appels aux relations de voisinage. Pour chaque relation appelée par l'utilisateur, le temps de réponse est le temps nécessaire à *ARVis* pour extraire le sous-ensemble de règles avec l'algorithme d'extraction locale correspondant, et pour générer le paysage (le coût en temps de cette dernière opération est négligeable devant la première). Les seuils d'indices de règle choisis dans les scénarios sont donnés dans le tableau 6.2. Pour ces expérimentations, le serveur applicatif d'*ARVis* était un SGI Origin 2000 équipé de quatre processeurs RISC R10000 à 250 MHz et de 512 Mo de mémoire.

Le premier jeu de données est la base MUSHROOMS de l'UCI Repository [BM98]. Il s'agit de données peu volumineuses mais qui sont connues pour être très corrélées. Le scénario d'exploration réalisé sur ce jeu de données est détaillé dans le tableau 6.3. Les deux autres jeux de données sont de grandes bases synthétiques, T10.I4.D100k et T20.I6.D100k. Elles ont été créées à l'aide du générateur[2] de données synthétiques d'IBM décrit dans [AS94a]. Le jeu de données T20.I6.D100k est délibérément très dense (en moyenne, chaque individu vérifie 43% des items). Les scénarios d'exploration pour ces deux jeux de données sont similaires à celui décrit dans le tableau 6.3. Ils ne sont pas détaillés puisque les données n'ont pas de signification réelle.

Comme le montre la figure 6.1, les temps de réponse tendent à décroître au fur et à mesure de l'exploration. Ceci est dû au dispositif de sauvegarde progressive des itemsets (pour les expérimentations, les tables contenant les itemsets et leurs cardinaux étaient vides au début des scénarios). Localement, des augmentations dans les temps de réponse (par exemple à t=6 et t=11 dans le scénario de la base MUSHROOMS) apparaissent quand les algorithmes d'extraction locale ont besoin d'une grande quantité d'itemsets qui n'ont pas encore été comptés.

[2] http ://www.almaden.ibm.com/software/quest/Resources/index.shtml

	MUSHROOMS	T10.I4.D100k	T20.I6.D100k
minSupport	1%	0.05%	0.05%
maxSupport	100%	100%	100%
minConf	70%	70%	70%
maxConf	100%	100%	100%
minLift	0	0	0
maxLift	100	100	100
minIPEE	0	0	0
maxIPEE	1	1	1
minII	0.5	0	0
maxII	1	1	1

TAB. 6.2 – Seuils utilisés dans les scénarios d'exploration

Dans ce cas, comme dans n'importe quelle procédure d'extraction d'itemsets fréquents (voir chapitre 3), l'algorithme doit requêter les données.

L'exploration réalisée sur le jeu de données T20.I6.D100k montre que *ARVis* peut traiter efficacement les données denses. En particulier, au cours de cette exploration, des règles très spécifiques contenant jusqu'à 15 items et présentant un support de 0.07% ont été découvertes. Du fait de l'explosion combinatoire, de telles règles ne pourraient jamais être extraites dans des données denses avec un algorithme exhaustif de type *Apriori*.

6.5 Conclusion

Dans ce chapitre, nous avons adapté l'extraction des règles au caractère interactif du post-traitement de règles dans la méthodologie *RF*. Pour cela, nous avons développé des algorithmes spécifiques pour l'extraction locale des règles. Ce sont des algorithmes qui exploitent les contraintes engendrées par les relations de voisinage, et sont spécialement optimisés pour ces contraintes. Ils permettent de n'extraire que les ensembles de règles que l'utilisateur souhaite visualiser, au fur et à mesure de son exploration. L'utilisateur joue ainsi le rôle d'une heuristique intégrée au sein de la procédure d'extraction de règles d'association. Grâce aux contraintes syntaxiques qui réduisent drastiquement l'espace de recherche, les algorithmes sont polynomiaux en fonction du nombre d'items. Cette approche locale donne la possibilité de s'affranchir des limites de l'extraction exhaustive (par exemple l'algorithme *Apriori*), c'est-à-dire principalement des temps d'exécution qui peuvent s'avérer rédhibitoires sur des données denses, quand l'explosion combinatoire ne rend pas l'extraction impossible. En particulier, même les règles très spécifiques peuvent être étudiées avec les algorithmes d'extraction locale.

Nous proposons également une stratégie plus perfectionnée pour l'extraction locale des règles : l'extraction *avec mémoire*, solution intermédiaire entre extrac-

tion exhaustive (mémoire pleine) et extraction sous contraintes (sans mémoire). Les algorithmes d'extraction avec mémoire ont l'avantage de sauvegarder les cardinaux des itemsets analysés pour pouvoir les réutiliser par la suite. Ceci permet d'éviter d'avoir à déterminer plusieurs fois le même cardinal. De plus, les algorithmes d'extraction avec mémoire tirent profit de leurs sauvegardes pour optimiser l'élagage de l'espace de recherche par les contraintes. Nos expérimentations montrent des temps de réponse qui tendent à diminuer au fur et à mesure que l'utilisateur explore les règles.

Temps	**Relation de voisinage**	**Règle de transition (sur laquelle la relation de voisinage est appliquée)**	**Nombre de règles générées**
t=1	*Chaînage avant*	$CLASS = edible \rightarrow GILL\_SIZE = broad$	3
t=2	*Chaînage avant*	$CLASS = edible, GILL\_SIZE = broad \rightarrow ODOR = none$	4
t=3	*Items communs*	$CLASS = edible, GILL\_SIZE = broad, ODOR = none \rightarrow STALK\_SHAPE = tapering$	3
t=4	*Généralisation de la prémisse*	$GILL\_SIZE = broad, ODOR = none, STALK\_SHAPE = tapering \rightarrow CLASS = edible$	3
t=5	*Prémisse commune*	$GILL\_SIZE = broad, STALK\_SHAPE = tapering \rightarrow ODOR = none$	7
t=6	*Conclusion commune*	$GILL\_SIZE = broad, STALK\_SHAPE = tapering \rightarrow RING\_NUMBER = one$	54
t=7	*Chaînage avant*	$CLASS = edible, GILL\_SIZE = broad, STALK\_SHAPE = tapering \rightarrow ODOR = none$	10
t=8	retour + *Généralisation de la prémisse*	$CLASS = edible, GILL\_SIZE = broad, STALK\_SHAPE = tapering \rightarrow ODOR = none$	3
t=9	minConf=60% + *Généralisation de la prémisse*	$GILL\_SIZE = broad, STALK\_SHAPE = tapering \rightarrow CLASS = edible$	2
t=10	*Chaînage avant*	$STALK\_SHAPE = tapering \rightarrow GILL\_SIZE = broad$	8
t=11	*Spécialisation d'exception*	$GILL\_SIZE = broad, STALK\_SHAPE = tapering \rightarrow CLASS = edible$	4
t=12	*Chaînage avant*	$GILL\_SIZE = broad \rightarrow CLASS = edible$	3
t=13	*Chaînage avant*	$CLASS = edible, GILL\_SIZE = broad \rightarrow ODOR = none$	4
t=14	*Chaînage avant*	$CLASS = edible, GILL\_SIZE = broad, ODOR = none \rightarrow STALK\_SHAPE = tapering$	10

Tab. 6.3 – Scénario d'exploration pour le jeu de données MUSHROOMS

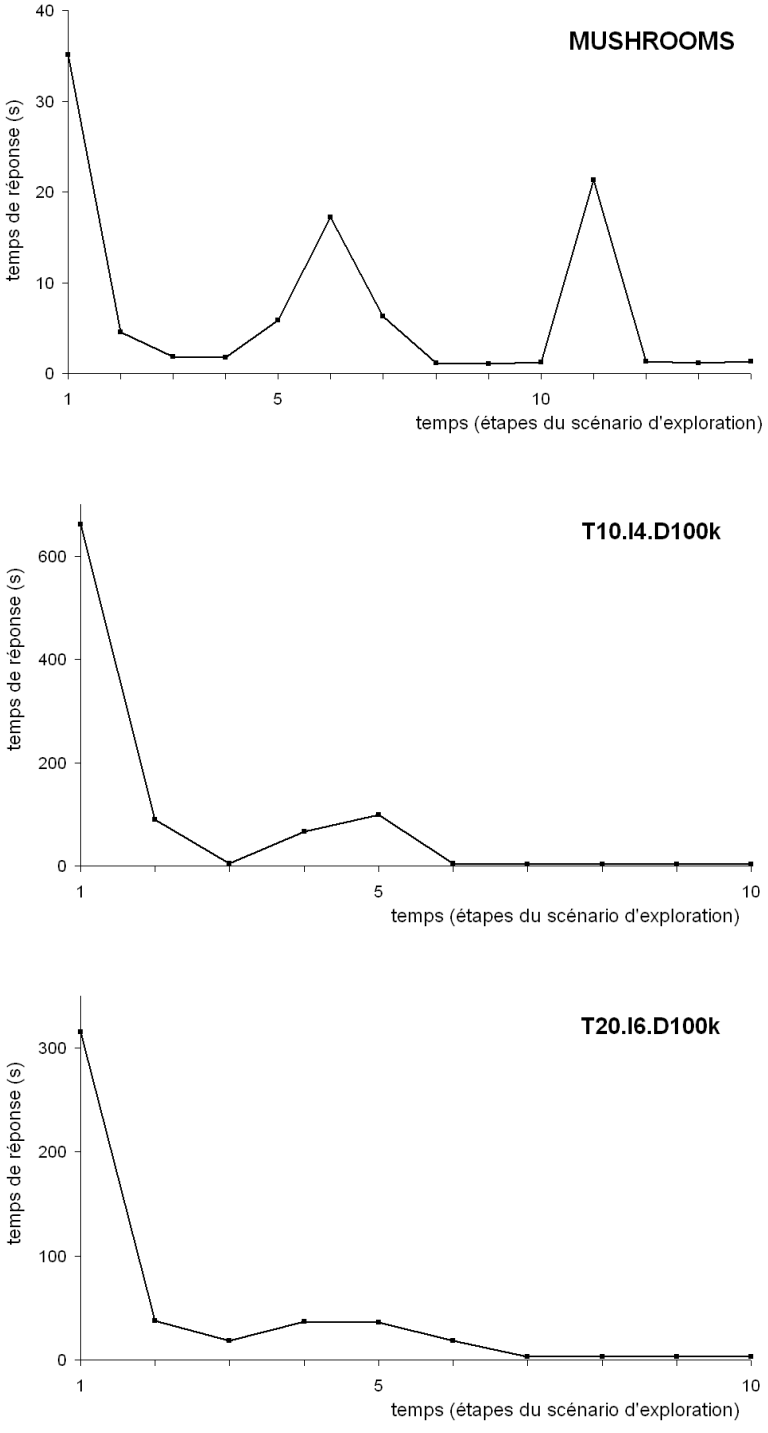


FIG. 6.1 – Temps de réponse obtenus sur trois scénarios d'exploration avec *ARVis*

Conclusion

Les algorithmes de fouille de données produisent des règles d'association en si grandes quantités que l'utilisateur ne peut généralement pas les exploiter directement. Pour identifier des connaissances intéressantes, il est nécessaire à la sortie des algorithmes de procéder à un post-traitement, consistant en une seconde opération de fouille. Alors que la fouille de données est réalisée automatiquement par des algorithmes combinatoires, la fouille de règles est laissée à la charge de l'utilisateur. Il s'agit d'une tâche laborieuse dans la pratique.

Les apports de notre approche

Les contributions de cette thèse peuvent être résumées ainsi :

- nous avons effectué une classification des indices de règle, et proposé de nouveaux indices ;
- nous avons établi une méthodologie pour la visualisation interactive des règles d'association ;
- nous avons développé des algorithmes spécifiques pour l'extraction locale des règles d'association ;
- nous avons réalisé un outil opérationnel pour la visualisation interactive des règles d'association.

La mesure de la qualité des règles

Notre première contribution est d'avoir établi un cadre formel pour l'étude des règles en définissant le concept de *règle*. Surtout, nous avons défini la notion d'*indice de règle* à la manière de Lerman avec les similarités [Ler81]. A notre connaissance, jamais une définition formelle qui puisse couvrir les nombreuses mesures de la littérature n'en avait été donnée. Nous avons également comparé les règles aux concepts connexes que sont les similarités, les implications, et les équivalences. Quand il s'agit de liaisons statistiques, les notions de similarité et d'équivalence sont souvent amalgamées.

Nous avons réalisé une classification inédite des indices de règle selon trois critères : l'objet, la portée, et la nature. Ces critères nous paraissent essentiels pour appréhender la signification des indices, et donc aussi pour aider l'utilisateur à choisir quels indices appliquer. La classification montre qu'il existe peu d'indices de règle au sens strict. Elle met également en évidence qu'aucun indice

statistique ne mesure l'écart à l'équilibre.

Enfin, nous avons proposé trois nouveaux indices de règle aux propriétés originales : l'indice probabiliste d'écart à l'équilibre IPEE, l'intensité d'implication entropique, et le taux informationnel. Ces trois indices ne sont pas concurrents : tandis que le taux informationnel est un indice descriptif, IPEE et l'intensité d'implication utilisés conjointement permettent de réaliser une évaluation statistique complète des règles. D'un point de vue plus pratique, le taux informationnel et l'intensité d'implication entropique sont des mesures perfectionnées qui combinent de multiples propriétés, utiles pour évaluer la qualité à l'aide d'une valeur unique. En revanche, IPEE est une mesure plus intelligible qui convient bien aux post-traitements de règles où divers indices sont associés.

La méthodologie *Rule Focusing*

Nous avons développé une méthodologie pour la visualisation interactive des règles d'association, nommée *Rule Focusing* (*RF*). Elle est conçue pour faciliter la tâche de l'utilisateur confronté à de grands ensembles de règles en prenant en compte ses capacités de traitement de l'information. Cette méthodologie centrée sur l'utilisateur permet de véritablement fouiller les ensembles de règles.

La méthodologie combine les trois principales approches qui sont traditionnellement proposées pour faciliter le post-traitement des règles : indices de règle, interactivité, et représentation visuelle. Elle consiste à laisser l'utilisateur explorer par lui-même des petits ensembles successifs de règles au moyen d'une visualisation interactive des règles et de leurs indices. En d'autres termes, l'utilisateur dirige une **suite d'explorations locales visuelles** en fonction de son intérêt pour les règles. Ainsi, un ensemble volumineux de règles est exploré sous-ensemble après sous-ensemble, de telle façon que l'utilisateur n'ait jamais à l'appréhender dans sa globalité. Cette approche est fondée sur :

- les principes cognitifs de traitement de l'information de Montgomery dans le contexte des modèles de décision [Mon83] ;
- les principes de visualisation d'information de Bertin pour la construction de représentations efficaces [Ber67].

Sur la base des principes cognitifs de Montgomery, nous développons le concept de *relation de voisinage* entre règles : des relations faisant sens pour l'utilisateur qui lui permettent d'isoler des sous-ensembles de règles limités et de naviguer entre les sous-ensembles. Ces relations de voisinage constituent une originalité forte de notre méthodologie en comparaison aux autres approches d'exploration de règles. Quant aux principes de visualisation de Bertin, nous les utilisons pour mettre en valeur les indices de règle et faciliter la reconnaissance des meilleures règles dans notre méthodologie.

Algorithmes pour l'extraction locale des règles

Nous avons proposé d'adapter l'extraction des règles au caractère interactif du post-traitement de règles dans la méthodologie *RF*. Pour cela, nous avons développé des algorithmes spécifiques pour l'extraction locale des règles. Ce sont des algorithmes à contraintes qui permettent de n'extraire que les ensembles

de règles que l'utilisateur souhaite visualiser, au fur et à mesure de son exploration. L'utilisateur joue ainsi le rôle d'une heuristique intégrée au sein de la procédure d'extraction de règles d'association. Grâce à des contraintes puissantes qui réduisent drastiquement l'espace de recherche, les algorithmes sont polynomiaux en fonction du nombre d'items. Cette approche locale donne la possibilité de s'affranchir des limites de l'extraction exhaustive (par exemple l'algorithme *Apriori*), c'est-à-dire principalement des temps d'exécution qui peuvent s'avérer rédhibitoires sur des données denses, quand l'explosion combinatoire ne rend pas l'extraction impossible. En particulier, même les règles très spécifiques (concernant une très faible portion des données) peuvent être étudiées avec les algorithmes d'extraction locale.

Nous proposons également une stratégie plus perfectionnée pour l'extraction locale des règles : l'extraction *avec mémoire*, solution intermédiaire entre extraction exhaustive (mémoire pleine) et extraction sous contraintes (sans mémoire). Les algorithmes d'extraction avec mémoire ont l'avantage de sauvegarder les résultats intermédiaires qu'ils produisent pour pouvoir les réutiliser. Ceci permet d'éviter d'avoir à générer plusieurs fois les mêmes résultats. De plus, les algorithmes d'extraction avec mémoire tirent profit de leurs sauvegardes pour optimiser l'élagage de l'espace de recherche par les contraintes. Nos expérimentations montrent des temps de réponse qui tendent à diminuer au fur et à mesure que l'utilisateur explore les règles.

L'outil de visualisation 3D *ARVis*

Nous avons développé un outil opérationnel pour la visualisation interactive des règles d'association, qui met en oeuvre les trois approches précédemment décrites. Cet outil nommé *ARVis* (*Association Rule Visualization*) permet d'explorer de grands volumes de règles et d'identifier les connaissances pertinentes.

ARVis repose sur une représentation 3D intuitive qui supporte de grands ensembles de règles décrits par plusieurs indices. Fondée sur la métaphore du paysage d'information, cette représentation inédite en visualisation de règles met en évidence les indices et facilite la reconnaissance des règles de bonne qualité. Pour ce qui concerne les interactions, plusieurs relations de voisinage réalisant des spécialisations ou généralisations permettent à l'utilisateur de guider sa navigation parmi les règles. Les tests d'*ARVis* sur données réelles montrent que l'outil aide à découvrir des règles intéressantes et de bonne qualité, et en particulier des règles localement dominantes (les meilleures dans la "région" de règles explorée).

Perspectives

Poursuivre la validation d'*ARVis* sur d'autres données

La validation d'*ARVis* peut être poursuivie selon deux axes. Tout d'abord, l'outil peut être testé avec un expert sur d'autres données (une campagne de tests sur des données de ressources humaines avec un expert de PerformanSe SA est prévue). Ceci permettrait de conforter nos hypothèses sur la stratégie

de fouille de règles de l'expert, mais aussi d'envisager de nouvelles relations de voisinage, soit sur conseil de l'expert (s'il est capable d'expliciter ces relations), soit par analyse des historiques d'exploration d'*ARVis*.

Ensuite, à la manière des expérimentations de Cockburn et McKenzie [CM01] ou Ware et Franck [WF96], il est possible de comparer *ARVis* à d'autres outils d'exploration de règles selon un protocole expérimental donné. Ce protocole peut consister à demander à deux panels d'utilisateurs d'utiliser chacun *ARVis* et un autre outil pour exécuter les mêmes tâches d'exploration sur les mêmes données, comme par exemple retrouver une règle précise ou répondre à une question portant sur les données. Une telle expérience permettrait de comparer l'exactitude et la rapidité des réponses des utilisateurs. Pour ce qui concerne plus particulièrement la représentation, les paysages d'*ARVis* peuvent être comparés aux autres types de représentations de règles en demandant aux utilisateurs d'identifier des règles en fonction de leurs items ou en fonction de leurs valeurs d'indices.

En outre, *ARVis* sera dans l'avenir utilisé en vision stéréoscopique immersive (avec lunettes stéréoscopiques et grand écran). Seule cette technologie permet en effet de tirer pleinement profit de la 3D. Le choix de la 3D plutôt que la 2D dans cette thèse avait d'ailleurs été largement motivé par la possibilité de visualisation en stéréoscopie. Cette utilisation d'*ARVis* sera une application de réalité virtuelle à part entière.

Approfondir la visualisation interactive des règles

L'exploration des règles peut être améliorée en développant de nouvelles relations de voisinage. En particulier, il serait sûrement utile d'intégrer dans les relations la possibilité d'utiliser une ou plusieurs hiérarchies d'items. Ceci permettrait de spécialiser ou généraliser les règles non pas uniquement en ajoutant ou supprimant des items, mais aussi en descendant ou en montant les items dans la hiérarchie. Un dispositif de mémorisation des règles pourrait également être mis en place dans *ARVis* pour donner à l'utilisateur la possibilité de rapidement retrouver, visualiser, et comparer les règles qu'il a jugées intéressantes lors de ses explorations passées. Plus généralement, les relations de voisinage créent un partitionnement des règles en sous-ensembles, mais rien ne garantit qu'il soit conforme au partitionnement mental que l'utilisateur effectue sur les règles qu'il explore (et ce malgré nos efforts pour le développement de relations qui font sens pour l'utilisateur). Ainsi, il serait bénéfique que l'utilisateur puisse créer ses propres paysages, en y plaçant les règles de son choix.

L'exploration des règles peut également être facilitée en procurant à l'utilisateur une "carte" de la navigation qu'il a effectuée. Cette carte pourrait prendre la forme par exemple d'un graphe indiquant les relations de voisinage et les règles de transition qui ont été choisies. Une telle représentation ferait office d'historique d'exploration, mais surtout permettrait, d'une certaine façon, de donner à l'utilisateur une vue globale de l'ensemble de règles qu'il visite. Etant donné que cet ensemble est trop volumineux pour être entièrement calculé, la vue globale ne serait pas générée a priori mais construite par l'utilisateur au fur et à mesure de sa navigation.

L'approche de visualisation interactive présentée dans cette thèse permet de véritablement naviguer dans la connaissance. Cependant, aujourd'hui seules les connaissances extraites des données sont prises en compte. L'approche pourrait être généralisée en autorisant l'introduction de connaissances extérieures, par exemple sous la forme d'ontologies du domaine étudié, de documents explicatifs des données, ou bien encore d'annotations écrites par l'utilisateur. Une telle association entre connaissances extraites des données et connaissances extérieures est ambitieuse, mais nous pensons que menée à bien, elle contribuerait à faciliter l'appropriation des connaissances par l'utilisateur.

Approfondir la mesure de la qualité des règles

Les indices de règle considèrent la prémisse et la conclusion d'une règle comme un tout, sans discerner les items qui les constituent. Pourtant, il serait intéressant de développer des mesures qui tiennent compte des liaisons entre chaque item en prémisse et chaque item en conclusion. D'une manière plus générale, il s'agit là d'un cas particulier du problème de la réduction des redondances entre règles : étant donnée une règle évaluée par un indice, la valeur de ce même indice sur une règle plus spécifique ou plus générale est-elle étonnante ou pas ? L'adaptation de indices de règle à la réduction des redondances mérite d'être étudiée de manière approfondie.

Un autre axe de recherche concernant les indices de règle consisterait à les étendre aux règles séquentielles. Les règles séquentielles sont une adaptation des règles d'association aux données séquentielles (par exemple temporelles) [MTV97] [Wei02]. Elles sont extraites par des algorithmes combinatoires, et le problème du volume de règles générées se pose également. Cependant, alors que les indices de règle sont nombreux, il existe très peu mesures de qualité de règles séquentielles. En particulier, aucune mesure n'évalue la significativité statistique de ces règles, alors même qu'une telle mesure permettrait de quantifier l'invraisemblance de la petitesse du nombre de contre-exemples eu égard à la longueur de la séquence étudiée, et aux fréquences d'occurrence de la prémisse et de la conclusion dans la séquence.

Bibliographie

[AAB+96] R. Agrawal, A. Arning, T. Bollinger, M. Mehta, J. Shafer & R. Srikant – « The quest data mining system », in *Proceedings of the second ACM SIGKDD international conference on knowledge discovery and data mining*, AAAI Press, 1996, p. 244–249.

[Agg02] C. C. Aggarwal – « Towards effective and interpretable data mining by visual interaction », *SIGKDD Explorations* **3** (2002), no. 2, p. 11–22.

[AIS93] R. Agrawal, T. Imielienski & A. Swami – « Mining association rules between sets of items in large databases », in *Proceedings of the 1993 ACM SIGMOD international conference on management of data*, ACM Press, 1993, p. 207–216.

[AMG+03] S. Aupetit, N. Monmarché, C. Guinot, G. Venturini & M. Slimane – « Exploration de données multimédia par réalité virtuelle », *Revue des Sciences et Technologies de l'Information* **17** (2003), no. 1-3, p. 71–82, Actes des journées Extraction et Gestion des Connaissances (EGC) 2003.

[AMS+96] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen & A. I. Verkamo – « Fast discovery of association rules », in *Advances in knowledge discovery and data mining*, American Association for Artificial Intelligence, 1996, p. 307–328.

[And95] K. Andrews – « Visualising cyberspace : information visualisation in the Harmony internet browser », in *Proceedings of the 1995 IEEE symposium on Information Visualization*, IEEE Computer Society, 1995, p. 97–104.

[AS94a] R. Agrawal & R. Srikant – « Fast algorithms for mining association rules », in *Proceedings of the twentieth international conference on very large data bases (VLDB 1994)* (J. B. Bocca, M. Jarke & C. Zaniolo, éds.), Morgan Kaufmann, 1994, p. 487–499.

[AS94b] C. Ahlberg & B. Shneiderman – « Visual information seeking : tight coupling of dynamic query filters with starfield displays », in *CHI'94 : Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM Press, 1994, p. 313–317.

[AY01] C. C. Aggarwal & P. S. Yu – « Mining associations with the collective strength approach », *IEEE Transactions on Knowledge and Data Engineering* **13** (2001), no. 6, p. 863–873.

[Aze03] J. Aze – « Une nouvelle mesure de qualité pour l'extraction de pépites de connaissances », *Revue des Sciences et Technologies de l'Information* **17** (2003), no. 1-3, p. 171–182, Actes des journées Extraction et Gestion des Connaissances (EGC) 2003.

[AZJ01] A. Ammoura, O. R. Zaiane & Y. Ji – « Immersed visual data mining : walking the walk », in *BNCOD 18 : Proceedings of the eighteenth British National Conference on Databases*, Springer-Verlag, 2001, p. 202–218.

[BA96] J. Brachman & T. Anand – « The process of knowledge discovery in databases : a human-centered approach », in *Advances in knowledge discovery and data mining* (U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth & R. Uthurusamy, éds.), AAAI/MIT Press, 1996, p. 37–58.

[BA99] R. J. Bayardo & R. Agrawal – « Mining the most interesting rules », in *Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining*, ACM Press, 1999, p. 145–154.

[Ban94] I. Bandhari – « Attribute focusing : machine-assisted knowledge discovery applied to software production process control », *Knowledge Acquisition journal* **6** (1994), no. 3, p. 271–294.

[Bay98] R. J. Bayardo – « Efficiently mining long patterns from databases », in *SIGMOD'98 : Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, ACM Press, 1998, p. 85–93.

[BBJ00] J.-F. Boulicaut, A. Bykowski & B. Jeudy – « Towards the tractable discovery of association rules with negations », in *Proceedings of the Fourth International Conference on Flexible Query Answering Systems (FQAS'00)* (H. L. Larsen et al., éds.), Advances in Soft Computing series, Physica-Verlag, 2000, p. 425–434.

[BBMM02] M. Botta, J.-F. Boulicaut, C. Masson & R. Meo – « A comparison between query languages for the extraction of association rules », in *Proceedings of the fourth international conference on data warehousing and knowlege discovery (DaWaK 2002)*, Lecture Notes in Computer Science, vol. 2454, Springer-Verlag, 2002, p. 1–10.

[BCKL02] D. Braga, A. Campi, M. Klemettinen & P. L. Lanzi – « Mining association rules from XML data », in *Proceedings of the fourth international conference on data warehousing and knowlege discovery (DaWaK 2002)*, Lecture Notes in Computer Science, vol. 2454, Springer-Verlag, 2002, p. 21–30.

[Bec97] B. G. Becker – « Research report : volume rendering for relational data », in *InfoVis'97 : Proceedings of the 1997 IEEE Symposium on Information Visualization*, IEEE Computer Society, 1997, p. 87–91.

[Ber67] J. Bertin – *Sémiologie graphique*, Gauthier-Villars, 1967, (3e édition en 1999 aux Editions de l'Ecole des Hautes Etudes en Sciences Sociales).

[BG+05] J. Blanchard, F. Guillet, , H. Briand & R. Gras – « Une version discriminante de l'Indice Probabiliste d'Ecart à l'Equilibre pour mesurer la qualité des règles », in *Actes de la troisième conférence Analyse Statistique Implicative (ASI2005)*, 2005, à paraître.

[BGB02] J. Blanchard, F. Guillet & H. Briand – « L'intensité d'implication entropique pour la recherche de règles de prédiction intéressantes dans les séquences de pannes d'ascenseurs », *Extraction des Connaissances et Apprentissage* **1** (2002), no. 4, p. 77–88, Actes des journées Extraction et Gestion des Connaissances (EGC) 2002.

[BGB03a] J. Blanchard, F. Guillet & H. Briand – « A virtual reality environment for knowledge mining », in *Proceedings of the fourteenth mini-EURO conference on Human Centered Processes HCP'2003*, 2003, p. 175–179.

[BGB03b] J. Blanchard, F. Guillet & H. Briand – « Exploratory visualization for association rule rummaging », in *Proceedings of the fourth international workshop on Multimedia Data Mining in conjunction with ACM KDD'2003*, 2003, p. 107–114.

[BGB03c] J. Blanchard, F. Guillet & H. Briand – « Une visualisation orientée qualité pour la fouille anthropocentrée de règles d'association », *Cahiers Romans de Sciences Cognitives* **1** (2003), no. 3, p. 79–100.

[BGB03d] J. Blanchard, F. Guillet & H. Briand – « A user-driven and quality-oriented visualization for mining association rules », in *Proceedings of the third IEEE international conference on data mining ICDM'03*, IEEE Computer Society, 2003, p. 493–496.

[BGBG05a] J. Blanchard, F. Guillet, H. Briand & R. Gras – « Assessing rule interestingness with a probabilistic measure of deviation from equilibrium », in *Proceedings of the eleventh international symposium on Applied Stochastic Models and Data Analysis ASMDA-2005*, ENST, 2005, p. 191–200.

[BGBG05b] J. Blanchard, F. Guillet, H. Briand & R. Gras – « IPEE : Indice Probabiliste d'Ecart à l'Equilibre pour l'évaluation de la qualité des règles », in *Actes de l'atelier Qualité des Données et des Connaissances en conjonction avec la conférence EGC'05*, 2005, p. 26–34.

[BGGB04] J. Blanchard, F. Guillet, R. Gras & H. Briand – « Mesurer la qualité des règles et de leurs contraposées avec le taux informationnel TIC », *Revue des Nouvelles Technologies de l'Information* **E-2** (2004), p. 287–298, Actes des journées Extraction et Gestion des Connaissances (EGC) 2004.

[BGGB05] J. Blanchard, F. Guillet, R. Gras & H. Briand – « Using information-theoretic measures to assess association rule interestingness », in *Proceedings of the fifth IEEE international conference on data mining ICDM'05*, IEEE Computer Society, 2005, à paraître.

[BGPK03] J. Blanchard, F. Guillet, F. Poulet & P. Kuntz – « Highly interactive data mining with virtual reality », in *Proceedings of the*

fifth Virtual Reality International Conference VRIC'2003, ISTIA Innovation, 2003, p. 221–228.

[BGRB03] J. Blanchard, F. Guillet, F. Rantière & H. Briand – « Vers une représentation graphique en réalité virtuelle pour la fouille interactive de règles d'association », *Revue des Sciences et Technologies de l'Information* **17** (2003), no. 1-3, p. 105–117, Actes des journées Extraction et Gestion des Connaissances (EGC) 2003.

[BKGG03] J. Blanchard, P. Kuntz, F. Guillet & R. Gras – « Implication intensity : from the basic statistical definition to the entropic version », in *Statistical Data Mining and Knowledge Discovery* (H. Bozdogan, éd.), Chapman and Hall/CRC Press, 2003, chapter 28, p. 473–485.

[BKGG04] J. Blanchard, P. Kuntz, F. Guillet & R. Gras – « Mesure de la qualité des règles d'association par l'intensité d'implication entropique », *Revue des Nouvelles Technologies de l'Information* **E-1** (2004), p. 33–43, numéro spécial Mesures de qualité pour la fouille de données.

[BL03] J. Blanchard & R. Lehn – « Des graphes à la réalité virtuelle pour l'extraction adaptative de règles d'association" », in *Actes de l'atelier Visualisation et Extraction Adaptative des Connaissances en conjonction avec la conférence EGC'03*, 2003.

[Bla68] N. M. Blachman – « The amount of information that *y gives about X* », *IEEE Transcations on Information Theory* **IT-14** (1968), no. 1, p. 27–31.

[BM92] J.-P. Barthélemy & E. Mullet – « A model of selection by aspects », *Acta Psychologica* **79** (1992), no. 1, p. 1–19.

[BM98] C. Blake & C. Merz – « UCI repository of machine learning databases », 1998, www.ics.uci.edu/˜mlearn/MLRepository.html.

[BMS97] S. Brin, R. Motwani & C. Silverstein – « Beyond market baskets : generalizing association rules to correlations », *SIGMOD Record* **26** (1997), no. 2, p. 265–276.

[BMUT97] S. Brin, R. Motwani, J. D. Ullman & S. Tsur – « Dynamic itemset counting and implication rules for market basket data », *SIGMOD Record* **26** (1997), no. 2, p. 255–264.

[BQK97] C. Brunk, J. Quelly & R. Kohavi – « Mineset : An integrated system for data mining », in *Proceedings of the third ACM SIGKDD international conference on knowledge discovery and data mining* (D. Heckerman, H. Mannila, D. Pregibon & R. Uthurusamy, éds.), AAAI Press, 1997, p. 135–138.

[Bra99] T. Bray – « Measuring the web », in *Readings in Information Visualization : Using vision to think* (S. Card, J. Mackinlay & B. Schneiderman, éds.), Morgan Kaufmann, 1999, p. 469–492.

[BSGG04] H. Briand, M. Sebag, R. Gras & F. Guillet (éds.) – *Mesures de qualité pour la fouille de données*, Cépaduès Editions, 2004, numéro spécial de la Revue des Nouvelles Technologies de l'Information.

[Bur01] D. Burdick – « MAFIA : a maximal frequent itemset algorithm for transactional databases », in *ICDE'01 : Proceedings of the seventeenth International Conference on Data Engineering*, IEEE Computer Society, 2001, p. 443.

[CB03] P. Coiffet & G. Burdea – *Virtual reality technology*, Wiley, 2003.

[CFB91] C. M. Carswell, S. Frankenberger & D. Bernhard – « Graphing in depth : perspectives on the use of three-dimensional graphs to represent lower-dimensional data », *Behaviour and Information Technology* **10** (1991), no. 6, p. 459–474.

[CG05] R. Couturier & R. Gras – « CHIC : traitement de données avec l'analyse implicative », *Revue des Nouvelles Technologies de l'Information* **E-3** (2005), p. 679–684, Actes des journées Extraction et Gestion des Connaissances (EGC) 2005.

[Che04] C. Chen – *Information visualization : beyond the horizon*, 2004.

[CM84] W. S. Cleveland & R. McGill – « Graphical perception : theory, experimentation, and application to the development of graphical methods », *Journal of the American Statistical Association* **79** (1984), no. 387, p. 531–554.

[CM01] A. Cockburn & B. McKenzie – « 3d or not 3d ? evaluating the effect of the third dimension in a document management system », in *CHI'01 : Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM Press, 2001, p. 434–441.

[CMS99] S. K. Card, J. Mackinlay & B. Shneiderman – *Readings in information visualization : Using vision to think*, Morgan Kaufmann, 1999.

[CN89] P. Clark & T. Niblett – « The CN2 induction algorithm », *Machine Learning* **3** (1989), no. 4, p. 261–283.

[Coh60] J. Cohen – « A coefficient of agreement for nominal scales », *Educational and Psychological Measurement* (1960), no. 20, p. 37–46.

[CRC03] A. Ceglar, J. F. Roddick & P. Calder – « Guiding knowledge discovery through interactive data mining », in *Managing data mining technologies in organizations : techniques and applications*, Idea Group Publishing, 2003, p. 45–87.

[CSBD97] E. F. Churchill, D. Snowdon, S. Benford & P. Dhanda – « Using VR-VIBE : browsing and searching for documents in 3D-space », in *Proceedings of the Seventh International Conference on Human-Computer Interaction (HCI International'97)* (M. J. Smith, G. Salvendy & R. J. Koubek, éds.), Elsevier, 1997, p. 857–860.

[Dic45] L. Dice – « Measures of the amount of ecologic association between species », *Ecology* (1945), no. 26, p. 297–302.

[Dun86] J. M. Dunn – « Relevance logic and entailment », in *Handbook of Philosophical Logic* (D. Gabbay & F. Guenthner, éds.), vol. 3, Kluwer Academic Publishers, 1986, p. 117–224.

[Ead84] P. Eades – « A heuristic for graph drawing », *Congressus Numerantium* **42** (1984), p. 149–160.

[EP96] J. F. Elder & D. Pregibon – « A statistical perspective on knowledge discovery in databases », in *Advances in knowledge discovery and data mining* (U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth & R. Uthurusamy, éds.), AAAI/MIT Press, 1996, p. 83–113.

[FGW01] U. Fayyad, G. Grinstein & A. Wierse (éds.) – *Information visualization in data mining and knowledge discovery*, Morgan Kaufmann, 2001.

[Fle96] L. Fleury – « Extraction de connaissances dans une base de données pour la gestion des ressources humaines », Thèse de doctorat, Université de Nantes, 1996.

[FMMT01] T. Fukuda, Y. Morimoto, S. Morishita & T. Tokuyama – « Data mining with optimized two-dimensional association rules », *ACM Transactions on Database Systems* **26** (2001), no. 2, p. 179–213.

[FMP01] P. Fuchs, G. Moreau & J. Papin – *Le traité de la réalité virtuelle*, Les presses de l'école des Mines de Paris, 2001.

[FPSM91] W. J. Frawley, G. Piatetsky-Shapiro & C. J. Matheus – « Knowledge discovery in databases : an overview », in *Knowledge Discovery in Databases* (G. Piatetsky-Shapiro & W. J. Frawley, éds.), AAAI/MIT Press, 1991, p. 1–30.

[FPSS96] U. M. Fayyad, G. Piatetsky-Shapiro & P. Smyth – « From data mining to knowledge discovery : an overview », in *Advances in knowledge discovery and data mining* (U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth & R. Uthurusamy, éds.), AAAI/MIT Press, 1996, p. 1–34.

[FR04] P. Fule & J. F. Roddick – « Experiences in building a tool for navigating association rule result sets », in *CRPIT'04 : Proceedings of the second Australasian workshop on information security, data mining, web intelligence, and software internationalisation* (J. Hogan, P. Montague, M. Purvis & C. Steketee, éds.), Australian Computer Society, Inc., 2004, p. 103–108.

[Fre98] A. A. Freitas – « On objective measures of rule surprisingness », in *Proceedings of the second European conference on principles of data mining and knowledge discovery (PKDD'98)* (J. Zytkow & M. Quafafou, éds.), Lecture Notes in Artificial Intelligence, vol. 1510, Springer-Verlag, 1998, p. 1–9.

[Fre99] A. A. Freitas – « On rule interestingness measures », *Knowledge-Based Systems Journal* **12** (1999), no. 5-6, p. 309–315.

[Fur86] G. W. Furnas – « Generalized fisheye views », in *CHI'86 : Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM Press, 1986, p. 16–23.

[Gan91] J.-G. Ganascia – « Charade : apprentissage de bases de connaissances », in *Induction symbolique et numérique à partir de données*

(Y. Kodratoff & E. Diday, éds.), Cépaduès Editions, 1991, p. 309–326.

[GCB+04] R. Gras, R. Couturier, J. Blanchard, H. Briand, P. Kuntz & P. Peter – « Quelques critères pour une mesure de qualité de règles d'association », *Revue des Nouvelles Technologies de l'Information* **E-1** (2004), p. 3–31, numéro spécial Mesures de qualité pour la fouille de données.

[GD86] J. Guigues & V. Duquenne – « Familles minimales d'implications informatives résultant d'un tableau de données binaires », *Mathématiques et Sciences Humaines* **24** (1986), no. 95, p. 5–18.

[GdB99] B. Goethals & J. V. den Bussche – « A priori versus a posteriori filtering of association rules », in *Proceedings of the SIGMOD'99 Workshop on Research Issues on Data Mining and Knowledge Discovery* (K. Shim & R. Srikant, éds.), 1999.

[GGGP02] R. Gras, F. Guillet, R. Gras & J. Philippé – « Réduction des colonnes d'un tableau de données par quasi-équivalence entre variables », *Extraction des Connaissances et Apprentissage* **1** (2002), no. 4, p. 197–202, Actes des journées Extraction et Gestion des Connaissances (EGC) 2002.

[Gib79] J. J. Gibson – *The ecological approach to visual perception*, 1979.

[GKB01] R. Gras, P. Kuntz & H. Briand – « Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données », *Mathématiques et Sciences Humaines* **39** (2001), no. 154-155, p. 9–29.

[GKCG01] R. Gras, P. Kuntz, R. Couturier & F. Guillet – « Une version entropique de l'intensité d'implication pour les corpus volumineux », *Extraction des Connaissances et Apprentissage* **1** (2001), no. 1-2, p. 69–80, Actes des journées Extraction et Gestion des Connaissances (EGC) 2001.

[GLW00] G. Grahne, L. V. S. Lakshmanan & X. Wang – « Efficient mining of constrained correlated sets », in *Proceedings of the sixteenth international conference on data engineering (ICDE 2000)*, IEEE Computer Society, 2000, p. 512–521.

[Gra96] R. Gras – *L'implication statistique : nouvelle méthode exploratoire de données*, La Pensée Sauvage Editions, 1996.

[Gui99] J.-M. Guinnebault – « Caractérisation des logiques nonmonotones et logique conditionnelle du deuxième ordre », Thèse de doctorat, Université de Rennes 1, 1999.

[Gui00] S. Guillaume – « Traitement des données volumineuses, mesures et algorithmes d'extraction de règles d'association et règles ordinales », Thèse de doctorat, Université de Nantes, 2000.

[Gui04] F. Guillet – *Mesures de la qualité des connaissances en ECD*, 2004, Tutoriel des journées Extraction et Gestion des Connaissances (EGC) 2004, www.isima.fr/~egc2004/Cours/Tutoriel-EGC2004.pdf.

[GZ01] K. Gouda & M. J. Zaki – « Efficiently mining maximal frequent itemsets », in *ICDM'01 : Proceedings of the 2001 IEEE International Conference on Data Mining*, IEEE Computer Society, 2001, p. 163–170.

[GZ03] B. Goethals & M. J. Zaki (éds.) – *FIMI'03, proceedings of the icdm 2003 workshop on frequent itemset mining implementations*, CEUR Workshop Proceedings, vol. 90, 2003.

[HAC00] J. Han, A. An & N. Cercone – « Cviz : an interactive visualization system for rule induction », in *AI'00 : Proceedings of the thirteenth Biennial Conference of the Canadian Society on Computational Studies of Intelligence*, Springer-Verlag, 2000, p. 214–226.

[Han98] J. Han – « Towards on-line analytical mining in large databases », *SIGMOD Record* **27** (1998), no. 1, p. 97–107.

[HBL01] M. Hascoët & M. Beaudouin-Lafon – « Visualisation interactive d'information », *Information - Interaction - Intelligence (I3)* **1** (2001), no. 1, p. 77–108.

[HDH+] M. C. Hao, U. Dayal, M. Hsu, T. Sprenger & M. H. Gross – « Visualization of directed associations in e-commerce transaction data », in *Proceedings of VisSym 2001*, p. 185–192.

[HF95] J. Han & Y. Fu – « Discovery of multiple-level association rules from large databases », in *VLDB'95 : Proceedings of the twenty-first International Conference on Very Large Data Bases*, Morgan Kaufmann Publishers Inc., 1995, p. 420–431.

[HFW+96] J. Han, Y. Fu, W. Wang, K. Koperski & O. Zaiane – « DMQL : a data mining query language for relational databases », in *Proceedings of the 1996 SIGMOD workshop on research issues on data mining and knowledge discovery (DMKD)*, 1996.

[HG02] J. Hipp & U. Güntzer – « Is pushing constraints deeply into the mining algorithms really what we want ? an alternative approach for association rule mining », *SIGKDD Explorations* **4** (2002), no. 1, p. 50–55.

[HGN00] J. Hipp, U. Güntzer & G. Nakhaeizadeh – « Algorithms for association rule mining a general survey and comparison », *SIGKDD Explorations* **2** (2000), no. 1, p. 58–64.

[HHC03] J. Han, X. Hu & N. Cercone – « A visualization model of interactive knowledge discovery systems and its implementations », *Information Visualization* **2** (2003), no. 2, p. 105–125.

[HHNT86] J. Holland, K. Holyoak, R. Nisbett & P. Thagard – *Induction : Processes of inference, learning and discovery*, MIT Press, 1986.

[HK00] J. Han & M. Kamber – *Data mining : concepts and techniques*, 2000.

[HLSL00] F. Hussain, H. Liu, E. Suzuki & H. Lu – « Exception rule mining with a relative interestingness measure », in *Proceedings of the fourth Pacific-Asia conference on knowledge discovery and data mining (PAKDD2000)*, Lecture Notes in Computer Science, vol. 1805, Springer-Verlag, 2000, p. 86–97.

[HMS01] D. J. Hand, H. Mannila & P. Smyth – *Principles of data mining*, 2001.

[HPY00] J. Han, J. Pei & Y. Yin – « Mining frequent patterns without candidate generation », in *SIGMOD'00 : Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, ACM Press, 2000, p. 1–12.

[HW01] H. Hofmann & A. Wilhelm – « Visual comparison of association rules », *Computational Statistics* **16** (2001), no. 3, p. 399–415.

[IM96] T. Imielinski & H. Mannila – « A database perspective on knowledge discovery », *Communications of the ACM* **39** (1996), no. 11, p. 58–64.

[IV99] T. Imielinski & A. Virmani – « MSQL : a query language for database mining », *Data Mining and Knowledge Discovery* **3** (1999), no. 4, p. 373–408.

[Jac01] P. Jaccard – « Etude comparative de la distribution florale dans une portion des Alpes et du Jura », *Bulletin de la Société Vaudoise des Sciences Naturelles* (1901), no. 37, p. 547–579.

[JB02] B. Jeudy & J.-F. Boulicaut – « Optimization of association rule mining queries », *Intelligent Data Analysis* **6** (2002), no. 4, p. 341–357.

[JL01] A. Johnson & J. Leigh – « Tele-immersive collaboration in the CAVE research network », in *Collaborative Virtual Environments : digital places and spaces for interaction* (E. Churchill, D. Snowdon & A. Munro, éds.), Springer-Verlag, 2001, p. 225–243.

[JS01] S. Jaroszewicz & D. A. Simovici – « A general measure of rule interestingness », in *Proceedings of the fifth European conference on principles of data mining and knowledge discovery (PKDD'01)*, Springer-Verlag, 2001, p. 253–265.

[Kei02] D. A. Keim – « Information visualization and visual data mining », *IEEE Transactions on Visualization and Computer Graphics* **8** (2002), no. 1, p. 1–8.

[KHC97] M. Kamber, J. Han & J. Chiang – « Metarule-guided mining of multi-dimensional association rules using data cubes », in *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*, 1997, p. 207–210.

[KMR+94] M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen & A. I. Verkamo – « Finding interesting rules from large sets of discovered association rules », in *Proceedings of the third international conference on information and knowledge management (CIKM 1994)*, ACM Press, 1994, p. 401–407.

[KMT96] M. Klemettinen, H. Mannila & H. Toivonen – « Interactive exploration of discovered knowledge : a methodology for interaction and usability studies », Tech. report, University of Helsinki, 1996, TR C-1996-3.

[Kod97] Y. Kodratoff – « L'extraction de connaissances à partir de données : un nouveau sujet pour la recherche scientifique », *Revue Électronique sur l'Apprentissage par les Données* **1** (1997), no. 1.

[Kod00] Y. Kodratoff – « Extraction de connaissances à partir des données et des textes », in *Actes des journées sur la fouille dans les données par la méthode d'analyse statistique implicative*, Presses de l'Université de Rennes 1, 2000, p. 151–165.

[Kro96] U. Krohn – « Vineta : navigation through virtual information spaces », in *AVI'96 : Proceedings of the workshop on Advanced Visual Interfaces*, ACM Press, 1996, p. 49–58.

[KT01] I. Kopanakis & B. Theodoulidis – « Visual data mining and modeling techniques », in *Proceedings of the KDD-2001 workshop on visual data mining*, 2001.

[Kul27] S. Kulczynski – « Die pflanzenassoziationen der pieninen », *Bulletin International de l'Académie Polonaise des Sciences et des Lettres. Classe des Sciences Mathématiques et Naturelles* (1927), no. suppl. II, p. 57–203, série B.

[KvdWW01] E. Kleiberg, H. van de Wetering & J. J. V. Wijk – « Botanical visualization of huge hierarchies », in *INFOVIS'01 : Proceedings of the IEEE Symposium on Information Visualization 2001 (INFOVIS'01)*, IEEE Computer Society, 2001, p. 87–94.

[LA04] I. C. Lerman & J. Azé – « Indice probabiliste discriminant de vraissemblance du lien pour des données volumineuses », *Revue des Nouvelles Technologies de l'Information* **E-1** (2004), p. 69–94, numéro spécial Mesures de qualité pour la fouille de données.

[Leh00] R. Lehn – « Un système interactif de visualisation et de fouille de règles pour l'extraction de connaissances dans les bases de données », Thèse de doctorat, Université de Nantes, 2000.

[Ler81] I. C. Lerman – *Classification et analyse ordinale des données*, Dunod, 1981.

[Ler91] I. C. Lerman – « Foundations in the likelihood linkage analysis classification method », *Applied Stochastic Models and Data Analysis* **7** (1991), p. 69–76.

[LFZ99] N. Lavrac, P. A. Flach & B. Zupan – « Rule evaluation measures : a unifying view », in *ILP'99 : Proceedings of the ninth International Workshop on Inductive Logic Programming*, Springer-Verlag, 1999, p. 174–185.

[LH96] B. Liu & W. Hsu – « Post-analysis of learned rules », in *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)*, AAAI Press, 1996, p. 828–834.

[LHCM00] B. Liu, W. Hsu, S. Chen & Y. Ma – « Analyzing the subjective interestingness of association rules », *IEEE Intelligent Systems* **15** (2000), no. 5, p. 47–55.

[LHM99] B. Liu, W. Hsu & Y. Ma – « Pruning and summarizing the discovered associations », in *KDD'99 : Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining*, ACM Press, 1999, p. 125–134.

[LHWC99] B. Liu, W. Hsu, K. Wang & S. Chen – « Visually aided exploration of interesting association rules », in *PAKDD'99 : Proceedings of the Third Pacific-Asia Conference on Methodologies*

for Knowledge Discovery and Data Mining, Springer-Verlag, 1999, p. 380–389.

[LLN02] C. K.-S. Leung, L. V. S. Lakshmanan & R. T. Ng – « Exploiting succinct constraints using FP-trees », *SIGKDD Explorations Newsletter* **4** (2002), no. 1, p. 40–49.

[LMV+04] P. Lenca, P. Meyer, B. Vaillant, P. Picouet & S. Lallich – « Evaluation et analyse multicritère des mesures de qualité des règles d'association », *Revue des Nouvelles Technologies de l'Information* **E-1** (2004), p. 219–246, numéro spécial Mesures de qualité pour la fouille de données.

[Loe47] J. Loevinger – « A systematic approach to the construction and evaluation of tests of ability », *Psychological Monographs* **61** (1947), no. 4.

[LRP95] J. Lamping, R. Rao & P. Pirolli – « A focus+context technique based on hyperbolic geometry for visualizing large hierarchies », in *CHI'95 : Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM Press/Addison-Wesley Publishing Co., 1995, p. 401–408.

[LT04] S. Lallich & O. Teytaud – « Evaluation et validation de l'intérêt des règles d'association », *Revue des Nouvelles Technologies de l'Information* **E-1** (2004), p. 193–218, numéro spécial Mesures de qualité pour la fouille de données.

[LV03] P. Lyman & H. R. Varian – « How much information ? », Tech. report, School of Information Management and Systems at the University of California at Berkeley, 2003, www.sims.berkeley.edu/how-much-info-2003.

[Mac95] A. M. MacEachren – *How maps work : representation, visualization, and design*, 1995.

[MLW00] Y. Ma, B. Liu & C. K. Wong – « Web for data mining : organizing and interpreting the discovered rules using the web », *SIGKDD Explorations* **2** (2000), no. 1, p. 16–23.

[Mon83] H. Montgomery – « Decision rules and the search for a dominance structure : towards a process model of decision making », in *Analysing and aiding decision processes* (P. Humphreys, O. Svenson & A. Vari, éds.), Elsevier Science Publishers, 1983, p. 343–369.

[Mos68] F. Mosteller – « Association and estimation in contingency tables », *Journal of the American Statistical Association* **63** (1968), no. 321, p. 1–28.

[MPC98] R. Meo, G. Psaila & S. Ceri – « An extension to SQL for mining association rules », *Data Mining and Knowledge Discovery* **2** (1998), no. 2, p. 195–224.

[MTV94] H. Mannila, H. Toivonen & A. I. Verkamo – « Efficient algorithms for discovering association rules », in *AAAI Workshop on Knowledge Discovery in Databases (KDD-94)* (U. M. Fayyad & R. Uthurusamy, éds.), AAAI Press, 1994, p. 181–192.

[MTV97] H. Mannila, H. Toivonen & A. I. Verkamo – « Discovery of frequent episodes in event sequences », *Data Mining and Knowledge Discovery* **1** (1997), no. 3, p. 259–289.

[Mun00] T. Munzner – « Interactive visualization of large graphs and networks », Thèse, Stanford University, 2000.

[NCJK01] A. A. Nanavati, K. P. Chitrapura, S. Joshi & R. Krishnapuram – « Mining generalised disjunctive association rules », in *CIKM'01 : Proceedings of the tenth international conference on Information and knowledge management*, ACM Press, 2001, p. 482–489.

[NLHP98] R. T. Ng, L. V. S. Lakshmanan, J. Han & A. Pang – « Exploratory mining and pruning optimizations of constrained associations rules », in *Proceedings of the 1998 ACM SIGMOD international conference on management of data* (L. M. Haas & A. Tiwary, éds.), ACM Press, 1998, p. 13–24.

[NVGB03] H. R. Nagel, M. Vittrup, E. Granum & S. Bovbjerg – « Exploring non-linear data relationships in VR using the 3D visual data mining system », in *Proceedings of the Third International Workshop on Visual Data Mining, in conjunction with the Third IEEE International Conference on Data Mining*, 2003.

[Och57] A. Ochiai – « Zoogeographic studies on the soleoid fishes found in japan and its neighbouring regions », *Bulletin of the Japanese Society of Scientific Fisheries* (1957), no. 22, p. 526–530.

[Pea96] K. Pearson – « Mathematical contributions to the theory of evolution : regression, heredity and panmixia », *Philosophical Transactions of the Royal Society Of London* **series A** (1896), no. 187, p. 253–318.

[PH00] J. Pei & J. Han – « Can we push more constraints into frequent pattern mining ? », in *KDD 2000 : Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM Press, 2000, p. 350–354.

[PS91] G. Piatetsky-Shapiro – « Discovery, analysis, and presentation of strong rules », in *Knowledge Discovery in Databases* (G. Piatetsky-Shapiro & W. J. Frawley, éds.), AAAI/MIT Press, 1991, p. 229–248.

[PT99] B. Padmanabhan & A. Tuzhilin – « Unexpectedness as a measure of interestingness in knowledge discovery », *Decision Support Systems* **27** (1999), no. 3, p. 303–318.

[Rae02] L. D. Raedt – « A perspective on inductive databases », *SIGKDD Explorations* **4** (2002), no. 2, p. 69–77.

[RCL+98] G. Robertson, M. Czerwinski, K. Larson, D. C. Robbins, D. Thiel & M. van Dantzich – « Data mountain : using spatial memory for document management », in *UIST'98 : Proceedings of the eleventh annual ACM symposium on user interface software and technology*, ACM Press, 1998, p. 153–162.

[RMC91] G. G. Robertson, J. D. Mackinlay & S. K. Card – « Cone trees : animated 3D visualizations of hierarchical information »,

in *Proceedings of the Conference on Human Factors in Computing Systems CHI'91*, ACM Press, 1991, p. 189–194.

[RR40] P. Russel & T. Rao – « On habitat and association of species of anopheline larvae in south-eastern madras », *Journal of the Malaria Institute of India* (1940), no. 3, p. 153–178.

[RR00] C. P. Rainsford & J. F. Roddick – « Visualisation of temporal interval association rules », in *Proceedings of the second international conference on intelligent data engineering and automated learning (IDEAL 2000)*, Springer-Verlag, 2000, p. 91–96.

[RT60] D. Rogers & T. Tanimoto – « A computer program for classifying plants », *Science* (1960), no. 132, p. 1115–1118.

[RZN01] G. Ritschard, D. A. Zighed & N. Nicoloyannis – « Maximisation de l'association par regroupement de lignes ou de colonnes d'un tableau croisé », *Mathématiques et Sciences Humaines* **39** (2001), no. 154-155, p. 81–97.

[SA95] R. Srikant & R. Agrawal – « Mining generalized association rules », in *VLDB'95 : Proceedings of the twenty-first International Conference on Very Large Data Bases*, Morgan Kaufmann Publishers Inc., 1995, p. 407–419.

[SA96a] R. Srikant & R. Agrawal – « Mining quantitative association rules in large relational tables », *SIGMOD Records* **25** (1996), no. 2, p. 1–12.

[SA96b] R. Srikant & R. Agrawal – « Mining sequential patterns : generalizations and performance improvements », in *EDBT'96 : Proceedings of the fifth International Conference on Extending Database Technology*, Springer-Verlag, 1996, p. 3–17.

[Sap90] G. Saporta – *Probabilités, analyse des données, et statistique*, Editions Technip, 1990.

[SB92] M. Sarkar & M. H. Brown – « Graphical fisheye views of graphs », in *CHI'92 : Proceedings of the SIGCHI conference on human factors in computing systems*, ACM Press, 1992, p. 83–91.

[SBM98] C. Silverstein, S. Brin & R. Motwani – « Beyond market baskets : generalizing association rules to dependence rules », *Data Mining and Knowledge Discovery* **2** (1998), no. 1, p. 39–68.

[SCK+97] J. Symanzik, D. Cook, B. D. Kohlmeyer, U. Lechner & C. Cruz-Neira – « Dynamic statistical graphics in the C2 virtual reality environment », *Computing Science and Statistics* **29** (1997), no. 2, p. 41–47.

[SG91] P. Smyth & R. M. Goodman – « Rule induction using information theory », in *Knowledge Discovery in Databases* (G. Piatetsky-Shapiro & W. J. Frawley, éds.), AAAI/MIT Press, 1991, p. 159–176.

[SG92] P. Smyth & R. M. Goodman – « An information theoretic approach to rule induction from databases », *IEEE Transactions on Knowledge and Data Engineering* **4** (1992), no. 4, p. 301–316.

[Shn96] B. Shneiderman – « The eyes have it : a task by data type taxonomy for information visualization », in *Proceedings of IEEE Symposium on Visual Languages VL'96*, IEEE Computer Society, 1996, p. 336–343.

[Shn02] B. Shneiderman – « Inventing discovery tools : combining information visualization with data mining », *Information Visualization* **1** (2002), no. 1, p. 5–12.

[Sim79] H. Simon – *Models of thought*, Yale University Press, 1979.

[SM58] R. Sokal & C. Michener – « A statistical method for evaluating systematic relationships », *University of Kansas Science Bulletin* (1958), no. 38, p. 1409–1438.

[SON95] A. Savasere, E. Omiecinski & S. Navathe – « An efficient algorithm for mining association rules in large databases », in *VLDB'95 : Proceedings of the twenty-first International Conference on Very Large Data Bases*, Morgan Kaufmann Publishers Inc., 1995, p. 432–444.

[SON98] A. Savasere, E. Omiecinski & S. Navathe – « Mining for strong negative associations in a large database of customer transactions », in *ICDE'98 : Proceedings of the Fourteenth International Conference on Data Engineering*, IEEE Computer Society, 1998, p. 494–502.

[Spe90] I. Spence – « Visual psychophysics of simple graphical elements », *Journal of Experimental Psychology : Human Perception and Performance* **16** (1990), no. 4, p. 683–692.

[Spe00] R. Spence – *Information visualization*, Addison Wesley, 2000.

[SS88] M. Sebag & M. Schoenauer – « Generation of rules with certainty and confidence factors from incomplete and incoherent learning bases », in *Proceedings of the European knowledge acquisition workshop EKAW'88*, Gesellschaft für Mathematik und Datenverarbeitung mbH, 1988, p. 28.1–28.20.

[ST96a] A. Silberschatz & A. Tuzhilin – « User-assisted knowledge discovery : how much should the user be involved », in *Proceedings of the 1996 SIGMOD workshop on research issues on data mining and knowledge discovery (DMKD)*, 1996.

[ST96b] A. Silberschatz & A. Tuzhilin – « What makes patterns interesting in knowledge discovery systems », *IEEE Transactions on Knowledge and Data Engineering* **8** (1996), no. 6, p. 970–974.

[SVA97] R. Srikant, Q. Vu & R. Agrawal – « Mining association rules with item constraints », in *Proceedings of the third ACM SIGKDD international conference on knowledge discovery and data mining* (D. Heckerman, H. Mannila, D. Pregibon & R. Uthurusamy, éds.), AAAI Press, 1997, p. 67–73.

[SW49] C. Shannon & W. Weaver – *The mathematical theory of communication*, University of Illinois Press, 1949.

[SZ05] E. Suzuki & J. M. Zytkow – « Unified algorithm for undirected discovery of exception rules », *International Journal of Intelligent Systems* **20** (2005), no. 7, p. 673–691.

[TA02] A. Tuzhilin & G. Adomavicius – « Handling very large numbers of association rules in the analysis of microarray data », in *KDD'02 : Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining*, ACM Press, 2002, p. 396–404.

[The70] H. Theil – « On the estimation of relationships involving qualitative variables », *American Journal of Sociology* **76** (1970), p. 103–154.

[Tis01] J. Tisseau – « Réalité virtuelle – autonomie *in virtuo* », mémoire d'Habilitation à Diriger des Recherches, Université de Rennes 1, 2001.

[TK00] P.-N. Tan & V. Kumar – « Interestingness measures for association patterns : a perspective », in *Proceedings of the KDD-2000 workshop on postprocessing in machine learning and data mining*, 2000.

[TKS04] P.-N. Tan, V. Kumar & J. Srivastava – « Selecting the right objective measure for association analysis », *Information Systems* **29** (2004), no. 4, p. 293–313.

[TNHB00] T. Teusan, G. Nachouki & J. P. Henri Briand – « Discovering association rules in large, dense databases », in *Proceedings of the fourth European Conference on Principles of Data Mining and Knowledge Discovery (PKDD-2000)*, Springer-Verlag, 2000, p. 638–645.

[Tuf83] E. Tufte – *The visual display of quantitative information*, Graphics Press, 1983.

[VLL04] B. Vaillant, P. Lenca & S. Lallich – « Etude expérimentale de mesures de qualité de règles d'association », *Revue des Nouvelles Technologies de l'Information* **E-2** (2004), p. 341–352, Actes des journées Extraction et Gestion des Connaissances (EGC) 2004.

[War00] C. Ware – *Information visualization : perception for design*, 2000.

[Wei02] G. M. Weiss – « Predicting telecommunication equipment failures from sequences of network alarms », in *Handbook of data mining and knowledge discovery*, Oxford University Press, Inc., 2002, p. 891–896.

[WF96] C. Ware & G. Franck – « Evaluating stereo and motion cues for visualizing information nets in three dimensions », *ACM Transactions on Graphics* **15** (1996), no. 2, p. 121–140.

[Wil96] G. J. Wills – « 288 ways to say "this is interesting" », in *INFOVIS'96 : Proceedings of the 1996 IEEE Symposium on Information Visualization*, IEEE Computer Society, 1996, p. 54–61.

[Wil05] L. Wilkinson – *The grammar of graphics*, 2005.

[WS92] C. Williamson & B. Shneiderman – « The dynamic homefinder : evaluating dynamic queries in a real-estate information exploration system », in *SIGIR'92 : Proceedings of the fifteenth annual international ACM SIGIR conference on Research and development in information retrieval*, ACM Press, 1992, p. 338–346.

[WS02] E. J. Wegman & J. Symanzik – « Immersive projection technology for visual data mining », *Journal of Computational and Graphical Statistics* **11** (2002), no. 1, p. 163–188.

[WWT99] P. C. Wong, P. Whitney & J. Thomas – « Visualizing association rules for text mining », in *Proceedings of the 1999 IEEE symposium on information visualization*, IEEE Computer Society, 1999, p. 120–123.

[WZZ04] X. Wu, C. Zhang & S. Zhang – « Efficient mining of both positive and negative association rules », *ACM Transactions on Information Systems* **22** (2004), no. 3, p. 381–405.

[Yul00] G. Yule – « On the association of attributes in statistics », *Philosophical Transactions of the Royal Society of London* **series A** (1900), no. 194, p. 257–319.

[Zak01] M. J. Zaki – « SPADE : an efficient algorithm for mining frequent sequences », *Machine Learning* **42** (2001), no. 1-2, p. 31–60.

[ZR00] D. A. Zighed & R. Rakotomalala – *Graphes d'induction*, Hermes Science Publications, 2000.

[ZZ96] R. Zembowicz & J. M. Zytkow – « From contingency tables to various forms of knowledge in databases », in *Advances in knowledge discovery and data mining*, American Association for Artificial Intelligence, 1996, p. 328–349.

Résumé :
De nombreuses méthodes d'Extraction de Connaissances dans les Données (ECD) produisent des résultats sous forme de règles. Les règles ont l'avantage de représenter les connaissances de manière explicite, ce qui en fait des modèles tout à fait intelligibles pour un utilisateur. Elles sont d'ailleurs au fondement de la plupart des théories de représentation de la connaissance en sciences cognitives. En fouille de données, la principale technique à base de règles est l'extraction de règles d'association, qui a donné lieu à de nombreux travaux de recherche.
La limite majeure des algorithmes d'extraction de règles d'association est qu'ils produisent communément de grandes quantités de règles, dont beaucoup se révèlent même sans aucun intérêt pour l'utilisateur. Ceci s'explique par la nature non supervisée de ces algorithmes : ne considérant aucune variable endogène, ils envisagent dans les règles toutes les combinaisons possibles de variables. Dans la pratique, l'utilisateur ne peut pas exploiter les résultats tels quels directement à la sortie des algorithmes. Un post-traitement consistant en une seconde opération de fouille se révèle indispensable pour valider les volumes de règles et découvrir des connaissances utiles. Cependant, alors que la fouille de données est effectuée automatiquement par des algorithmes combinatoires, la fouille de règles est une tâche laborieuse à la charge de l'utilisateur.
La thèse développe deux approches pour assister l'utilisateur dans le post-traitement des règles d'association :
– la mesure de la qualité des règles par des indices numériques,
– la supervision du post-traitement par une visualisation interactive.
Pour ce qui concerne la première approche, nous formalisons la notion d'indice de qualité de règles et réalisons une classification inédite des nombreux indices de la littérature, permettant d'aider l'utilisateur à choisir les indices pertinents pour son besoin. Nous présentons également trois nouveaux indices aux propriétés originales : l'indice probabiliste d'écart à l'équilibre, l'intensité d'implication entropique, et le taux informationnel. Pour ce qui concerne la seconde approche, nous proposons une méthodologie de visualisation pour l'exploration interactive des règles. Elle est conçue pour faciliter la tâche de l'utilisateur confronté à de grands ensembles de règles en prenant en compte ses capacités de traitement de l'information. Dans cette méthodologie, l'utilisateur dirige la découverte de connaissances par des opérateurs de navigation adaptés en visualisant des ensembles successifs de règles décrits par des indices de qualité.
Les deux approches sont intégrées au sein de l'outil de visualisation *ARVis* (*Association Rule Visualization*) pour l'exploration interactive des règles d'association. *ARVis* implémente notre méthodologie au moyen d'une représentation 3D, inédite en visualisation de règles, mettant en valeur les indices de qualité. De plus, *ARVis* repose sur un algorithme spécifique d'extraction sous contraintes permettant de générer les règles interactivement au fur et à mesure de la navigation de l'utilisateur. Ainsi, en explorant les règles, l'utilisateur dirige à la fois l'extraction et le post-traitement des connaissances.

Mots-clés :
Extraction de Connaissances dans des bases de Données, visualisation d'information, fouille de connaissances, règles d'association, exploration des règles, visualisation interactive des règles, mesures de qualité de règles, extraction de règles sous contraintes

Abstract :
Numerous methods of Knowledge Discovery in Databases (KDD) produce results in the form of rules. Rules have the advantage of representing knowledge explicitly, which makes them absolutely intelligible models for a user. Besides, they are a major element of most theories of knowledge representation in cognitive sciences. In data mining, the main rule-based paradigm is association rules which have received significant research attention.
The main limit of association rule mining algorithms is that they commonly generate large amounts of rules, many of which do not even have any interest for the user. This is due to the unsupervised nature of these algorithms : as they consider no endogenous variable, they search for all the possible combinations of variables in the rules. In practice, the user cannot exploit the results directly at the output of the algorithms. A post-process consisting in a second analysis is indispensable to validate the sets of rules and discover useful knowledge. However, whereas data analysis is automatically computed by combinatorial algorithms, rule analysis is a tedious task manually done by the user.
The thesis develops two approaches for assisting the user in association rule post-processing :
– assessing rule interestingness with numerical indexes,
– supervising the post-process with an interactive visualization.
With regard to the first approach, we define the concept of rule interestingness measure and present a novel classification of the numerous measures of the literature, helping the user to choose the measures relevant to his/her application. We also describe three new measures with original properties : the probabilistic index of deviation from equilibrium, the entropic implication intensity, and the directed information ratio. As for the second approach, we propose a visualization methodology for the interactive exploration of rules. It facilitates the task of the user faced with large sets of rules by taking into account his/her information processing abilities. In this methodology, the user drives the knowledge discovery with appropriate navigation operators by visualizing successive sets of rules described by interestingness measures.
The two approaches are integrated into the visualization tool *ARVis* (Association Rule Visualization) for interactive exploration of association rules. *ARVis* implements our methodology by means of a 3D representation -the only one of its kind in rule visualization-which highlights the interestingness measures. Moreover, *ARVis* relies on a specific constraint-based rule-mining algorithm allowing to generate the rules interactively throughout the user navigation. Thus, by exploring the rules, the user drives both the knowledge extraction and the knowledge post-processing.

Keywords :
Knowledge Discovery in Databases, information visualization, knowledge post-processing, association rules, rule exploration, interactive rule visualization, rule interestingness measures, constraint-based rule-mining