N° d'ordre : 2007- ? ?- ? ? ? Année 2007

THÈSE

présentée devant

L'Institut National des Sciences Appliquées de Lyon

pour obtenir

Le Grade de Docteur

Spécialité

Informatique

Ecole Doctorale : Informatique et Information pour la Société

par

Clément Fauré

Découvertes de motifs pertinents par l'implémentation d'un réseau bayésien : application à l'industrie aéronautique

Soutenue publiquement le XX/XX/2007 devant le jury :

Jean-François BOULICAUT, Professeur à l'INSA de Lyon	Co-directeur
Jean CHARLET, Chercheur HDR AP - Hopitaux de Paris	Rapporteur
Sylvie DEPRAT, Chercheur au CCR EADS	
Bart GOETHALS, Chercheur à l'Université d'Anvers (Belgique)	
François JACQUENET, Professeur à l'Université de Saint-Etienne	Rapporteur
Alain MILLE, Professeur à l'Université Lyon 1	Co-directeur

A tous, un grand Merci.

Résumé

Dans un contexte industriel un ingénieur est souvent confronté à analyser un ensemble de données relatives à un processus opérationnel. L'environnement dans lequel est plongé le modèle évoluant constamment au cours du temps, on va constater de manière inévitable l'apparition de différences entre ce qui était attendu et ce qui est réellement observé. Plus *inquiètant*, certains comportements peuvent être masqués dans la masse des données. Il faut alors être en mesure de déceler ces différences et, le cas échéant, de mettre à jour le modèle utilisé. Un apport combiné des techniques d'extraction de la connaissance (ECD) et de méthodes issues de l'ingénierie de la connaissance permet de répondre à ce besoin.

Dans cette thèse, nous avons envisagé la découverte de règles d'association pertinentes. A partir d'un ensemble de données on est capable d'extraire un ensemble de motifs décrivant les particularités – locales – des données. Cependant, l'étude de ces résultats d'extraction se révèle souvent laborieuse, de part la complexité des motifs manipulés et de part le manque d'outils qui permettraient de faciliter leur analyse.

Dans un premier temps nous avons étudié une généralisation des approches pour la génération de règles d'association non redondantes. Cela nous a permis de travailler à partir d'ensembles concis, ne contenant pas de redondance intrinsèque. Puis, nous avons proposé la mise en place d'un processus de découverte de connaissance qui intègre la définition, et l'exploitation d'un réseau bayésien pour faciliter l'analyse de règles extraites. L'évolution de ce modèle est facilitée par la découverte de règles pertinentes, elles mêmes rendues plus accessibles grâce à l'évolution du modèle. Nous avons également défini le rôle et l'importance de l'expert au sein de ce processus. Enfin, nous avons montré l'application de nos propositions au domaine des interruptions opérationnelles dans l'industrie aéronautiques.

Mots clés

Extraction de connaissances, ingénierie de la connaissace, règles d'associaiton, réseaux bayésiens.

Table des matières

Table des figures

Liste des tableaux

Introduction

Ce manuscrit présente nos travaux de recherche sur la découverte de règles d'association pertinentes par l'exploitation d'un Réseau Bayésien. Ces travaux sont appliqués à un cas d'application industriel qui concerne l'aide à l'analyse de données d'interruptions opérationnelles dans l'industrie aéronautique.

Contributions de la thèse, organisation du mémoire

Dans le cadre des travaux réalisés sur l'aide à l'analyse des données d'interruptions opérationnelles, pour le compte d'un grand constructeur aéronautique, nous nous sommes particulièrement intéressés à l'élaboration d'une boucle *vertueuse* permettant la découverte de règles d'association intéressantes.

L'approche que nous proposons est basée sur :
- l'extraction d'une collection de règles non redondantes aux propriétés particulières,
- l'utilisation d'un Réseau Bayésien pour la modélisation des dépendances connues du domaine d'application,
- la définition et l'exploitation de mesures d'intérêt prenant en compte les connaissances du domaine,
- l'exploitation d'annotations réalisées par l'expert sur les règles d'association

Ces différents points sont regroupés au sein d'un processus itératif dont le principal objectif est d'arriver à faciliter la découverte de règles d'association intéressantes, mais aussi, par effet de bord, permettre la définition et la consolidation d'un Réseau Bayésien qui capture les principales dépendances du domaine.

Tout d'abord, le premier chapitre introduit le contexte industriel des travaux de la thèse, là savoir l'aide à l'analyse des données d'interruptions opérationnelles. On y présente aussi les différents niveaux de la problématique, aussi bien du point de vu industriel, qu'au niveau de l'ingénierie de la connaissance et de la fouille de données ; ainsi que les voies que nous avons décidé d'aborder.

Ensuite, le chapitre 2 est l'occasion de présenter le cadre des règles d'associa-

tion, les représentations condensées, ainsi que les techniques actuelles pour le post-traitement de la collection de règles extraites. On y détaille plus particulièrement l'approche qui a initié nos travaux de recherche, à savoir les travaux de S. Jaroszewicz [JS04]. Notre approche reposant sur la technique des Réseaux Bayésiens ce chapitre donne également l'opportunité d'introduire cette technique, utilisée ici pour capturer et exploiter les principales dépendances du domaine.

Le chapitre 3 décrit les travaux réalisés, il reprend les différents points de notre approche, et les principales contributions sur les deux axes qui sont : l'analyse de règles d'association et l'ingénierie de la connaissance. Ces travaux sont illustrés et validés sur un exemple expérimental issu de la littérature des réseaux bayésiens, le réseau *VisitAsia*.

Nous verrons ensuite, au chapitre 4, un application pratique de notre approche sur les données d'interruptions opérationnelles. Nous tirerons les conclusions quant aux limites actuelles rencontrées sur des données réelles.

Enfin le chapitre 5 présente une discussion dans laquelle nous replaçons nos contributions dans le cadre de la découverte de règles d'association et de l'ingénierie de la connaissance, et nous dégageons quelques perspectives.

Chapitre 1

Cadre de travail, problématique de la recherche

Les travaux de thèse présentés dans ce rapport résultent de la collaboration entre l'équipe « Ingénierie et systèmes apprenants » du centre commun de recherche EADS et les départements « Fouille de données » et « Ingénierie de la connaissance » du LIRIS[1]. Dans ce premier chapitre nous présentons le contexte industriel qui a initié les travaux de thèse : l'aide à l'analyse des données d'interruptions opérationnelles pour le compte d'un grand constructeur aéronautique.

1.1 Le contexte industriel

Dans un contexte industriel, un ingénieur est souvent confronté à l'analyse de grands volumes de données, produites et stockées à des fins de test, de validation, ou encore dans le but de tracer le fonctionnement d'un processus opérationnel. Ces données peuvent notamment servir à faciliter la détection de comportements non prévus du système. Dans ce cas de figure, il s'agit de retrouver les particularités de fonctionnement qui diffèrent des modélisations initiales. En effet tout processus opérationnel étant soumis aux aléas du monde réel, il n'est pas rare que, malgré toutes les précautions prises lors des phases de conception du système, celui-ci diffère du comportement attendu.

Une des principales causes à l'origine de ce constat vient du décalage temporel qui existe entre la phase de conception initiale et l'exploitation en milieu opérationnel. L'environnement dans lequel est plongé le système évoluant constamment au cours du temps, on va constater de manière inévitable des différences entre ce qui était attendu

[1] http ://liris.cnrs.fr

et ce qui est observé. Il faut donc pouvoir détecter ces évolutions et, le cas échéant, les prendre en compte au sein du système.

Le contexte des interruptions opérationnelles illustre bien ce problème. Dans le domaine aéronautique, une interruption opérationnelle est un retard au départ (décollage) de plus de quinze minutes, une annulation ou une interruption de vol suite à un problème technique (panne ou dysfonctionnement). De tels événements sont aujourd'hui considérés avec un réelle importance par les compagnies aériennes du fait des coûts élevés qu'ils génèrent. Tout au long de ce document on utilisera l'abréviation IO pour désigner une *interruption opérationnelle*.

Les ingénieurs vont devoir analyser un ensemble de données relatives aux incidents afin d'en retirer un modèle de prédiction découlant du fonctionnement des appareils en service. Pour ce faire ils utilisent les connaissances non formalisées du domaine (rapports d'incidents, discussions, mels échangés, comptes-rendus de réunions), ainsi qu'un ensemble de méthodes propres à leur métier, qui ciblent des problèmes bien spécifiques. Par exemple, ils vont devoir mesurer l'impact du positionnement d'un équipement, et donc du temps additionnel nécessaire pour y accéder, sur les taux d'interruptions opérationnelles de l'appareil.

Une IO peut avoir des causes diverses. Certaines sont difficilement prévisibles ou évitables (comme des conditions météorologiques extrêmes par exemple), d'autres sont inhérentes au fonctionnement des compagnies aériennes (disponibilité des pièces et/ou des équipes de maintenance, décisions prises par le pilote), d'autres enfin sont directement imputables à l'avionneur (choix technologiques, fiabilité des équipements utilisés, positionnement de certains équipements pour une maintenance plus rapide, redondance des systèmes embarqués, etc). Comme on peut s'en rendre compte, il existe de nombreux paramètres qui rendent le domaine de l'analyse des interruptions opérationnelles complexe.

L'avionneur s'intéresse évidemment à la part des IO qui lui sont imputées. Ainsi, il doit pouvoir anticiper ces événements en donnant aux compagnies aériennes une estimation des performances opérationnelles de leurs appareils. Lors du lancement de nouveaux projets avions, les ingénieurs doivent fournir dès la phase de conception une prédiction la plus réaliste possible de la fréquence des interruptions opérationnelles lors de la future exploitation commerciale des avions. Dans la pratique, une part des IO est directement liée aux choix de conception. De ce fait, les objectifs IO vont initier, guider et valider le processus de développement. Enfin, pour aider les experts à mesurer l'impact des décisions techniques prises en termes de taux d'IO, un outil informatique a été mis en place. Il implémente un modèle mathématique stochastique intégrant les paramètres dont les impacts sur la fréquence des IO sont connus. Cet outil est calibré et paramétré par le retour d'expérience obtenu à partir d'avions, de systèmes ou d'équipements en service comparables. La figure 1.1 présente sous la forme d'un diagramme d'activité UML, une vue simplifiée du processus lié à la

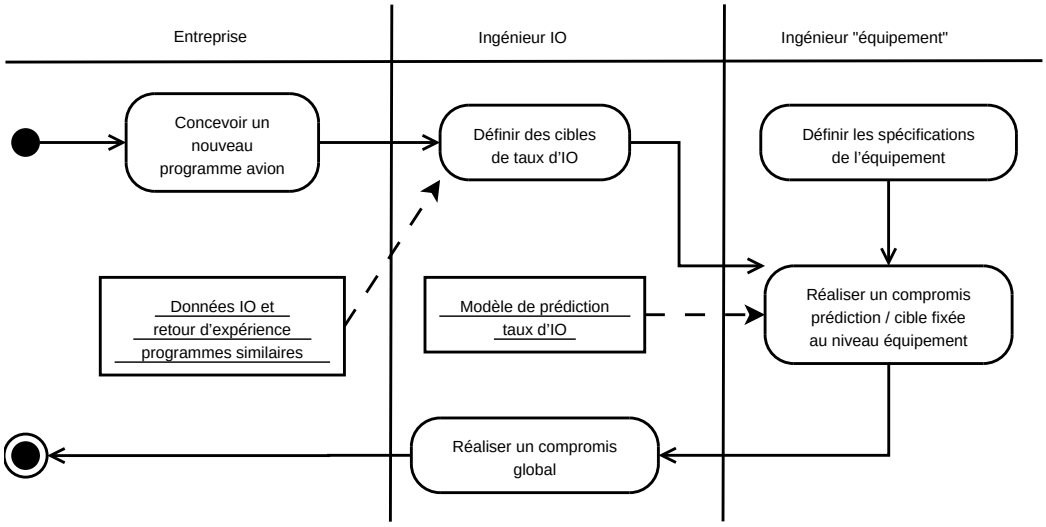

FIG. 1.1 – Diagramme de séquence simplifié présentant la problématique du cas d'application

modélisation des performances, en termes de taux d'IO, d'un nouveau programme avion.

Cette vue – simplifiée – du cas d'application nous permet d'introduire une des problématiques rencontrées par un industriel aéronautique. Celui-ci doit être en mesure d'exploiter le retour d'expérience sur des programmes passés pour prédire le taux d'interruptions opérationnelles d'un avion en phase de conception. Par *retour d'expérience*, on entend ici aussi bien l'ensemble des données produites par les avions en opération (détails des incidents, caractéristiques de l'avion, heures de vol, etc), que l'ensemble des informations dont disposent les ingénieurs, spécialistes des taux d'interruptions opérationnelles. L'objectif étant de développer le modèle de prédiction le plus précis possible.

1.2 Pratiques actuelles sur les données d'interruptions opérationnelles

Il faut bien faire la distinction entre les différentes problématiques qui interviennent autour de nos travaux de recherche :

1. d'une part la problématique des ingénieurs aéronautiques pour le problème décrit, c'est en quelque sorte le travail quotidien des experts chargés de l'étude et de la prédiction des IO,
2. d'autre part la contribution que l'on va apporter au cas d'application, c'est-à-

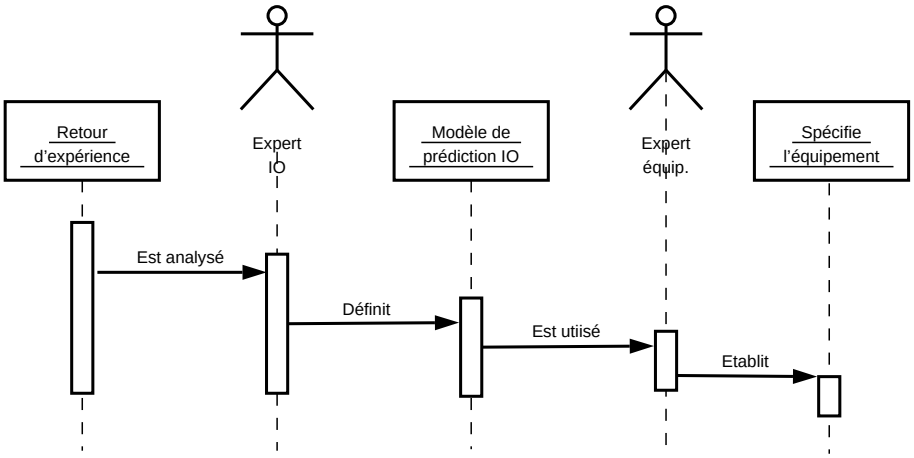

FIG. 1.2 – Diagramme de séquence simplifié présentant la problématique du cas d'application

dire la mise en œuvre d'un ensemble d'outils et de méthodes issus de la fouille de données et de l'ingénierie de la connaissance, afin de faciliter la découverte de facteurs contributeurs des taux d'IO,

3. et enfin les questions liées à la problématique de fouille de données que l'on se propose d'aborder.

Le premier axe, qui occupe particulièrement les ingénieurs spécialistes de la fiabilité opérationnelle, porte sur l'élaboration d'un modèle de prédiction de la fiabilité opérationnelle. Pour une application au domaine de l'aéronautique, le lecteur se reportera aux travaux de [HCC02]. L'approche adoptée par les auteurs est basée sur la résolution avec l'outil *Supercab* (développé par *Cab Innovation*) d'un processus de Markov périodique et borné. Les résultats obtenus par le biais de ce modèle montrent qu'il est possible de prédire l'impact des paramètres de conception sur les performances opérationnelles globales de l'avion. Cependant, pour que ce modèle soit le plus précis possible, il doit intégrer un grand nombre de paramètres issus des pratiques de maintenance et de la spécification des systèmes et des équipements.

L'axe de recherche qui nous concerne plus directement est l'identification des différents facteurs qui contribuent aux taux d'interruptions opérationnelles (et donc directement aux performances opérationnelles), à partir de grandes bases de données détaillant les incidents survenus en opération.

Actuellement il n'existe pas de travaux formalisés sur l'aide à l'analyse de données d'interruptions opérationnelles. Les outils employés par les ingénieurs sont des outils commerciaux (tableurs, systèmes de bases de données relationnelles) ainsi qu'un ensemble de méthodes *ad hoc* qui permettent de valider certains facteurs suggérés par l'expert. Il y a néanmoins un intérêt marqué quant à l'étude et la mise en place de

techniques permettant de faciliter cette phase d'analyse des données. Un des objectifs étant de pouvoir faire émerger les conditions opérationnelles pouvant entraîner de manière répétée des interruptions opérationnelles. Dans nos travaux nous nous sommes concentrés sur l'extraction de motifs locaux pertinents pour l'enrichissement des connaissances liées aux IO. On désigne par « motif local » une particularité observée sur les données. Ce motif peut s'exprimer sous différentes formes, notamment comme une association de facteurs co-occurants au sein d'un sous-ensemble des données étudiées.

Problématique des experts pour la prédiction des taux d'interruptions opérationnelles

Le premier point concerne le travail des ingénieurs aéronautiques chargés de définir et de maintenir le modèle de prédiction. Ce besoin concerne l'intégration des connaissances du domaine et du retour d'expérience pour l'amélioration du modèle de prédiction des taux d'interruptions opérationnelles. Concernant cette problématique, les besoins de recherche portent sur l'amélioration des modèles de calcul utilisés par le modèle de prédiction et notamment la détection et la validation de nouveaux facteurs contribuant aux IO. Certains de ces facteurs sont identifiés de manière informelle : à la suite de discussions, d'échanges de mails, ou de réunions entre les experts du domaine. D'autres ont été identifiés mais n'ont pas pu être intégrés au modèle de prédiction, faute de pouvoir les vérifier sur les données. D'autres enfin restent à découvrir à partir du retour d'expérience sur les programmes avions en service.

Une des voies, envisagée par les experts dans le domaine des IO pour la découverte de nouveaux facteurs d'interruptions, passe par une analyse poussée des données en service, la prise en compte des évolutions techniques, ainsi que l'acquisition de nouvelles connaissances du domaine et leurs intégrations au modèle de prédiction. Cependant, il n'y a actuellement pas de processus bien défini pour la recherche de ces nouvelles connaissances. L'expert émet un ensemble d'hypothèses qu'il va ensuite vérifier manuellement sur les données.

Un exemple concret est la découverte et la prise en compte d'un facteur ayant une grande influence sur les taux d'IO. L'intuition initiale de l'expert était de regarder dans les rapports de maintenance associés à l'ensemble des interruptions opérationnelles lesquels contenaient une référence à l'application d'une procédure de maintenance particulière (la *Master Minimum Equipment List* ou MMEL). Après une étude approfondie des données il s'est avéré que cette procédure avait effectivement une influence forte sur les taux d'IO. Cette analyse, conduite par les experts du domaine, a d'une part permis de quantifier l'importance de l'application de cette procédure à différents niveaux d'équipements, et, d'autre part, a poussé l'intégration de cet élément au modèle de prédiction.

Cet exemple montre bien qu'actuellement les méthodes utilisées peuvent être qualifiées d'*ad hoc*. En pratique, il s'agit par exemple de partir d'un export d'une base de données d'interruptions, puis d'effectuer un ensemble de statistiques et de requêtes manuelles pour confirmer ou non les hypothèses prises. La découverte de nouveaux éléments contributeurs des taux d'IO dépend donc presque exclusivement des intuitions formulées par l'expert et du nombre d'heures qu'il peut consacrer à leur vérification.

Problématique du cas d'application

Pour faciliter cette tâche d'analyse du retour d'expérience, on va différencier deux types de besoin :

- le premier est celui du *test d'hypothèse* qui doit permettre à l'expert de corroborer des propriétés sur les données collectées,
- le deuxième est celui de la *découverte de connaissances à partir des données*.

Dans ce contexte, la mise en place d'un processus d'extraction de connaissances à partir des données est potentiellement intéressante. L'application de ces techniques sur les données en service doit permettre la découverte de nouveaux facteurs qui pourraient être intégrés aux modèles de prédiction de la fréquence des IO. Plus précisément, on définit la *découverte d'une connaissance utile* comme étant un élément de connaissance qui, une fois présenté à l'expert par le biais des techniques de fouille, va faciliter la formalisation de nouveaux facteurs d'IO.

Pour ce faire on s'oriente vers l'utilisation de techniques issues de la fouille de données permettant de découvrir des associations de facteurs relatifs à des situations particulières d'IO. Cette proposition est présentée dans la figure 1.3.

Problématique de la découverte de règles d'association

Enfin, le troisième point concerne la contribution scientifique de ce travail de thèse. Telle qu'elle a été définie, la problématique industrielle nous porte à réfléchir sur des problèmes que l'on peut exprimer en des termes plus *génériques*. Ainsi on va se poser la question de la découverte de connaissances utiles à l'expert par l'exploitation du retour d'expérience et des connaissances du domaine. En particulier on va s'intéresser à l'instrumentation d'un processus ECD permettant de telles découvertes. Comment exploiter la connaissance du domaine pour faciliter la découverte de connaissances utiles ? Et inversement, comment intégrer les connaissances découvertes dans les modèles représentant la connaissance du domaine ? Quelles techniques mettre en place pour y arriver ?

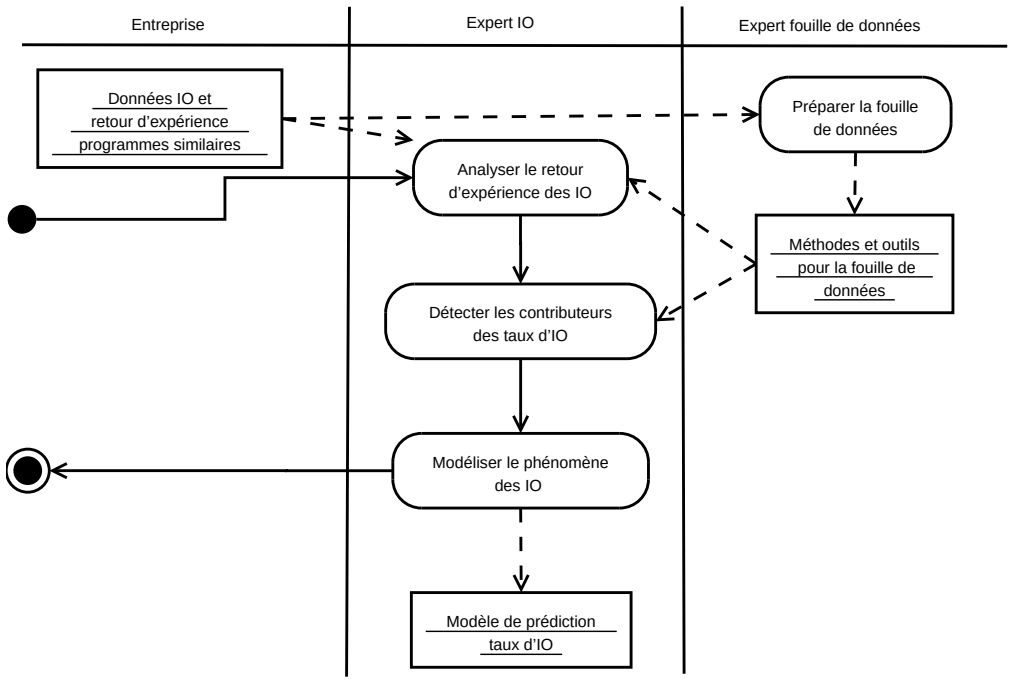

FIG. 1.3 – Présentation de l'approche envisagée pour la découverte de facteurs contribuant aux IO.

1.3 Découvrir des connaissances utiles à l'expert à partir du retour d'expérience

Nous avons introduit l'expression « découverte de connaissances utiles à l'expert ». Avant de continuer, il convient d'expliciter cette formule que l'on emploiera fréquemment tout au long de ce document. Plutôt que de donner une définition générale, on préfère faire le lien avec notre cas d'application. La définition que l'on propose se décompose en plusieurs parties puisqu'elle fait intervenir les termes de *connaissance* et *d'utilité*, ainsi que l'action *découvrir*.

1.3.1 Connaissance

Une connaissance peut être vue en tant qu'objet permettant l'*action*. C'est-à-dire qu'elle réunit à la fois une information et un mode d'emploi permettant d'utiliser cette information.

M. Polanyi [Pol66] a distingué deux types de connaissances : les connaissances tacites et les connaissances explicites. Les connaissances tacites sont les connaissances qui appartiennent au monde des objets et des représentations mentales. Elles re-

groupent les compétences innées ou acquises, le savoir-faire et l'expérience. Elles sont généralement difficiles à *formaliser*. Dans le cas de connaissances tacites, le mode d'emploi nécessaire pour utiliser une information est en quelque sorte *incorporé* chez l'homme.

Par exemple, un pilote d'avion qui doit prendre la décision de ne pas partir sous certaines conditions (état du train d'atterrissage, valve anti-gel à remplacer...) même si les conditions sont jugées acceptables par les équipes de maintenance au sol (et donc en accord avec les autorités de régulation). Dans ce cas de figure c'est l'expérience du pilote qui joue : celui-ci est au courant des faits (état de l'appareil, feu vert des équipes de maintenance) mais son propre mode d'emploi lui dicte la conduite à suivre.

Par opposition, les connaissances explicites sont les connaissances clairement articulées au niveau d'un document écrit, ou d'un système informatique. Ces connaissances sont transférables physiquement, car elles apparaissent sous une forme tangible (dossier papier ou électronique). Ici, le mode d'emploi est décrit de façon a pouvoir être exécuté.

Le cas typique de connaissance explicite est le manuel de recherche de pannes (TSM ou *troubleshooting manual*) destiné aux équipes de maintenance. Il s'agit véritablement d'un mode d'emploi sous forme de document écrit, qui va permettre à partir de son exécution de définir l'origine d'une panne.

Par ailleurs, il est aussi important de bien faire la distinction entre l'information, les données brutes, et la connaissance qui, elle, est l'appropriation et l'interprétation des informations par les hommes. Cette information elle-même contenue à l'état brut dans les données.

Dans les entreprises, la connaissance correspond au capital d'expertise que détiennent les hommes dans les différents domaines qui constituent le cœur de métier de l'entreprise. Dans notre contexte, on définit une connaissance comme étant un *fait* pouvant se vérifier sur les données issues d'un processus expérimental ou de données réelles.

Par exemple, un expert des données IO va savoir que la probabilité pour qu'il y ait une IO liée à l'équipement n°212042 est de $0,001\%$ lorsqu'il est embarqué sur un avion de type $A3X0$. Cette connaissance peut se vérifier sur les données en service du programme avion $A3X0$, et elle va pouvoir être utilisée, grâce aux connaissances de l'expert et aux outils qu'il a développés, pour estimer la probabilité d'IO de cet équipement embarqué sur un avion de type $A3Y0$.

1.3.2 Connaissance utile

Les motifs extraits vont revêtir des caractères différents aux yeux de l'expert. On s'intéresse plus particulièrement à la découverte de motifs que l'on qualifiera

d'« utile ». Cette *utilité* introduit la notion d'une plus value engendrée par la découverte de ce motif, par rapport à la compréhension initiale que l'on a sur le domaine. On peut la mesurer en fonction de la faculté que va avoir ce motif à être exploité, c'est-à-dire à être réinterprété dans le formalisme de l'expert et intégré aux modèles existants. On peut aussi évaluer l'utilité en fonction du nombre ou de l'importance des actions concernées par la découverte.

Une connaissance se révèle utile par rapport à un contexte donné. Le caractère d'utilité est donc intrinsèquement subjectif. Dans notre contexte, une connaissance est jugée utile si, potentiellement, elle permet (ou facilite) la découverte de nouveaux facteurs contribuant aux taux d'interruptions opérationnelles.

Si on reprend l'exemple utilisé précédemment concernant l'impact de l'application d'une procédure de maintenance particulière sur la fréquence des IO, on voit bien qu'il s'agit d'une connaissance utile : ce facteur a été intégré au modèle de prédiction de la fréquence des IO. Ainsi, non seulement il est pris en compte à chaque fois qu'on souhaite réaliser une estimation des IO, mais il a aussi un impact important sur la précision de l'estimation.

Dans le cadre de la fouille de données, on parlera de *motifs pertinents* pour désigner un motif qui a entraîné la découverte d'une connaissance utile pour l'expert.

1.3.3 Découverte de connaissances utiles à partir de données

L'action de *découvrir* une connaissance consiste à présenter de manière explicite ce qui était implicitement contenu dans les données. Il y a de nombreuses méthodes envisageables pour réaliser cette tâche. Dans nos travaux de thèse, on s'intéresse plus particulièrement au domaine de l'*extraction de connaissances à partir des données* (on utilisera l'acronyme ECD). Cette expression a été introduite pour la toute première fois dans [FPSM92], en tant que « processus d'extraction non triviale de connaissances implicites, non connues à l'avance et potentiellement intéressantes, à partir des données ».

Par rapport à notre problématique, la mise en place d'un processus ECD (présenté dans la figure 1.4) pose les objectifs suivants :

- être capable de formaliser, dans une certaine mesure, des savoirs spécifiques au domaine d'application de l'expert, savoirs souvent non formalisés tels les savoir-faire et procédures complexes résultant de l'expérience,
- fournir, par le biais de la fouille de données, les informations utiles, et seulement elles, avec un minimum d'intervention de la part de l'expert ;
- et enfin permettre de capitaliser les informations collectées grâce à la fouille de données, de manière organisée, afin de les pérenniser.

Classiquement le processus ECD fait ressortir trois étapes, à savoir :

1. La phase de **préparation des données**, consiste dans un premier temps à développer une bonne compréhension du domaine d'application, des connaissances pertinentes du domaine ainsi que des objectifs de l'utilisateur final. Ensuite il faut mettre en place le jeu de données : sélection des données, réduction du nombre de variables, nettoyage et pré-traitements en vue des algorithmes des fouilles, gestion des données manquantes, etc. Cette phase est généralement très coûteuse en temps.
2. La tâche d'**extraction de modèles et motifs**. Le choix de l'algorithme d'extraction reflète les objectifs de fouille qui ont été fixés. Est-ce que le processus de fouille a pour but la classification, la régression, le *clustering*, l'extraction de règles... ? Une fois la technique choisie, il faut décider quels paramètres sont les plus appropriés par rapport aux données étudiées. Ainsi on peut distinguer deux types d'approche pour la fouille de données : la construction de modèles des données et l'extraction de motifs locaux.
3. Enfin, **l'exploitation des résultats** consiste à interpréter et analyser les motifs ou modèles extraits lors de l'étape (2). Il s'agit de faire correspondre les résultats obtenus avec les objectifs initialement fixés par l'utilisateur du système, donc de réinterpréter les résultats de la fouille par rapport à la problématique afin d'en tirer une nouvelle connaissance. On peut aussi ajouter à cette étape la consolidation des découvertes et leur éventuelle intégration aux modèles utilisés par l'expert.

Le processus ECD est itératif : de nombreux cycles doivent être réalisés sur les étapes (1) et (2) avant de pouvoir obtenir des résultats exploitables dans la phase (3).

L'information extraite par les algorithmes de fouille pourra (1) soit être organisée par un expert du domaine sous forme de modèle de classification ou de prédiction, (2) soit être utilisée pour préciser la définition de modèles existants, (3) ou encore fournir une représentation synthétique des données étudiées. Dans [HMS01] les auteurs définissent les algorithmes de fouille de données de la manière suivante : « *un algorithme de fouille de données est une procédure bien définie qui prend des données en entrée et qui produit une sortie sous la forme de modèles ou de motifs* ». L'expression « bien définie » indique ici que la procédure de fouille de données doit se terminer en un temps raisonnable sur une échelle humaine.

1.4 Apprendre et acquérir, ou construire un modèle des connaissances

On a commencé à faire la distinction entre la fouille de données orientée *modèle* et la fouille de données pour la découverte de *motifs*. Cette section présente les principales différences entre ces deux types d'approches et argumente le choix qui a été fait dans

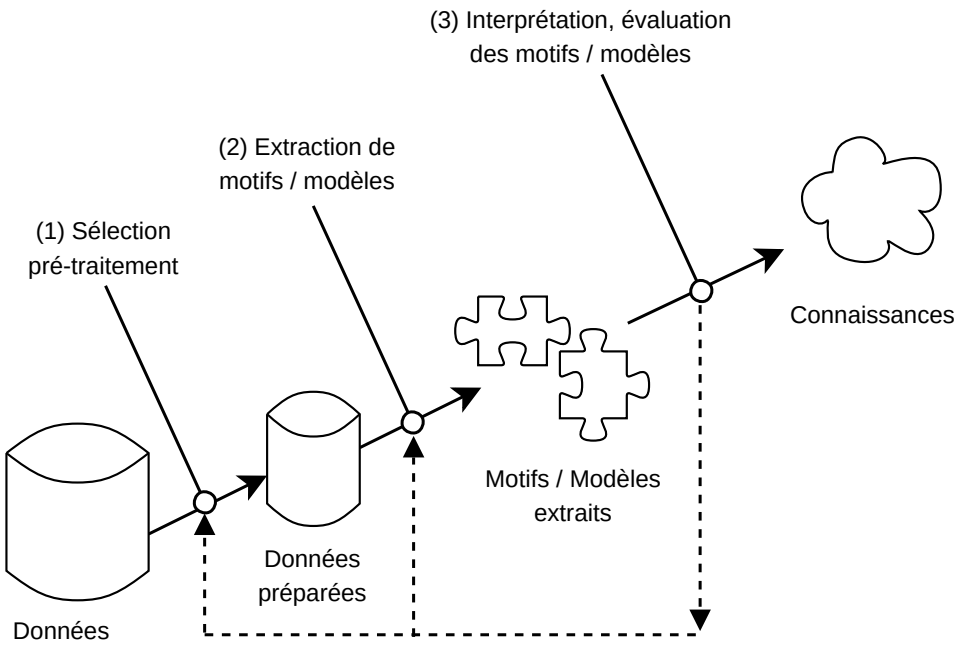

FIG. 1.4 – Processus simplifié d'Extraction de Connaissances à partir des Données

le cadre des travaux de thèse. Il faut garder à l'esprit que ces deux approches, bien que distinctes dans leur philosophie, ne s'excluent pas mutuellement dans la pratique.

Un *modèle*, tout d'abord, est une vision haut-niveau, une description générale du jeu de données sur lequel on travaille. Cette vision peut être descriptive (vue condensée et pratique sur les données), ou utilisée pour ses capacités d'inférence, c'est-à-dire que cette approche donne la possibilité à l'utilisateur de tirer des conclusions factuelles sur la population issue des données. Des exemples de modèles couramment employés sont les modèles de régression, les modèles de mélanges gaussiens, les Réseaux Bayésiens, etc.

Un *motif* quant à lui, décrit une propriété locale des données, qui ne se vérifie peut être que sur quelques individus (enregistrements) et/ou pour quelques variables. Il peut s'agir par exemple d'un point d'inflexion sur une courbe de régression, d'un ensemble d'éléments prenant des valeurs inhabituelles dans certaines conditions, etc. De même que pour les modèles, on peut rechercher des motifs pour leur aspect descriptif ou leur capacité d'inférence.

Pour illustrer ces deux approches, les auteurs de « *Principles of Data Mining* » [HMS01] proposent une analogie avec le domaine de la compression de données. Prenons un émetteur E qui doit envoyer une image (ou des données) I à un récepteur R. Il y a deux stratégies possibles : (a) envoyer toutes les données (*pixels* de l'image I) ou (b) envoyer une version compressée de cette image (un résumé en quelque sorte). La

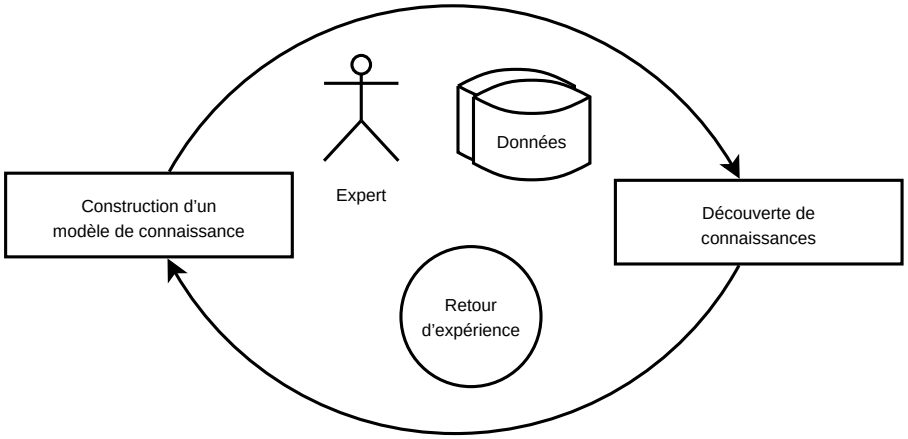

FIG. 1.5 – Collaboration des approches « modèles" et « motifs"

fouille de données, au sens large, correspond à la seconde approche : la compression est réalisée par le biais de techniques de fouille de données, soit en représentant les données d'origine en tant que modèle, soit (mais ce n'est pas exclusif) en identifiant les caractéristiques inhabituelles des données par le biais de motifs.

Prenons le cas d'application des interruptions opérationnelles. Un modèle est construit à partir du retour d'expérience et de l'expertise du domaine. Ce modèle permet de simuler le comportement général du système. Des motifs sont, quant à eux, un moyen de représenter des conditions inhabituelles qui entraînent des interruptions opérationnelles. Il est important de bien insister sur le fait que les approches à base de modèles ou à base de motifs ne sont pas en contradiction, ni en compétition. On se rend compte ici de l'importance de disposer de ces deux approches ; cela nous permet de nous poser les questions auxquelles on va répondre dans ces travaux de thèse : est-il possible d'utiliser le modèle pour faciliter la découverte de motifs ? Et réciproquement, la découverte de motifs peut elle contribuer à l'élaboration du modèle ?

Nous avons envisagé la collaboration de ces deux approches dans le cadre de la découverte de règles d'associations pertinentes.

1.5 La problématique de la découverte de règles d'association utiles à l'expert

1.5.1 Choix des règles d'association pour notre cas d'application

Il ne faut pas perdre de vue les différents objectifs que nous nous sommes fixés dans le cadre de ces travaux de thèse. En particulier, une des contributions industrielle

envisagée est de pouvoir faciliter la découverte de facteurs qui contribuent aux taux d'interruptions opérationnelles, à partir de grandes bases de données décrivant les retards des appareils en service.

Dans notre cas, il n'y a pas de *classes* à proprement parlé (tous les enregistrements représentent des interruptions opérationnelles), l'expert souhaite simplement voir apparaître des relations, entre les attributs de la base, caractéristiques de situations de retard.

On se situe donc dans un contexte *non supervisé* : le processus de fouille n'est pas nécessairement dirigé par le choix d'une classe d'attribut ou par des hypothèses prises par l'expert. De plus, la taille et la nature des données est aussi un facteur déterminant dans le choix des techniques utilisées.

Les règles d'association sont un exemple de motif local particulièrement étudié dans la littérature relative à la fouille de donnée. On reviendra plus en détails sur les arguments qui ont motivé leur utilisation : le postulat de départ étant que face à un problème de fouille de données non supervisée, il est très raisonnable de penser que les règles d'association peuvent être utilisées pour permettre la découverte de connaissances utiles à l'expert.

Les règles d'associations présentent l'avantage d'être facilement compréhensibles par un utilisateur ayant reçu un minimum de sensibilisation. Une fois générées, l'utilisateur peut donc être relativement autonome pour la phase d'analyse et de post-traitement des règles. De plus les algorithmes actuels nous autorisent à travailler sur des jeux de données arbitrairement grands, ce qui correspond bien aux pré-requis fixés par la problématique initiale.

On verra par la suite que les propriétés de ces règles vont nous permettre d'envisager leur utilisation pour faire évoluer – dans une certaine mesure – un modèle des connaissances du domaine. Ce problème se rapproche de l'ingénierie de la connaissance : il nous faudra donc étudier comment l'expert *considère* une règle d'association, et comment il envisage l'utiliser pour mettre à jour les connaissances du domaine.

1.5.2 Les règles d'association

Elles ont été initialement introduites par R. Agrawal [AIS93] en 1993. Ces règles sont généralement extraites à partir d'une matrice booléenne (ou base de données binaire). Une règle d'association, $A \rightarrow B$, peut se lire de la façon suivante : « Lorsque j'observe la présence des événements A dans les données, alors les événements B sont souvent observés ». On lui associe généralement des fonctions d'évaluation permettant en particulier de quantifier les termes « observe » et « souvent » (respectivement, la fréquence et la confiance), mais aussi de mesurer différents critères statistiques calculés à partir des données.

Le Tableau 1.1 montre un exemple de base de données binaire. Dans cet exemple, les colonnes représentent les différents attributs de la base, chaque ligne représente un enregistrement (ou transaction). Un « 1 » à l'intersection d'une ligne l et d'une colonne c indique que l'on observe la présence de l'attribut c pour l'enregistrement l. Par exemple, si les colonnes représentent différents produits d'un supermarché et chaque ligne le panier d'un client, alors les « 1 » permettent d'identifier le contenu de ces paniers. Ce tableau fait aussi apparaître deux exemples de règles d'association pouvant être extraites à partir de la matrice booléenne.

T_{id}	A	B	C	D	E
1	0	0	1	0	1
2	1	1	1	1	1
3	1	0	1	1	0
4	1	1	1	1	0
5	0	0	1	1	0
6	1	1	0	0	1

1. B $\rightarrow$ A (fréquence $= 0,5$, confiance $= 1,00$)
2. A B $\rightarrow$ C (fréquence $= 0,33$, confiance $= 0,66$)

TAB. 1.1 – Exemple de matrice booléenne et de règles d'association extraites.

Depuis que l'algorithme APRIORI a été proposé [AMS+96], une des principales préoccupations des différentes équipes de recherche a été l'amélioration des performances d'extraction. Plus précisément, il s'agissait d'être en mesure de calculer la collection d'itemsets fréquents sur des jeux de données arbitrairement grands, là où l'algorithme APRIORI montrait très vite ses limites. On verra notamment que ces réflexions ont abouti à la définition d'une représentation intermédiaire des données, dite *représentation condensée*. Ce type de représentation a été introduit la première fois en 1996 [MT96], puis de nombreuses propositions ont suivi. Le principe des représentations condensées est de calculer une représentation plus succincte des données initiales, tout en contrôlant (dans certains cas) la perte d'information. Ceci permet de réaliser le calcul de la fonction d'évaluation de manière plus efficace, et ouvre ainsi la porte à l'extraction de règles d'association sur des jeux de données de plus en plus denses, et faisant intervenir de nombreux attributs.

Pour nos travaux de thèse, nous avons retenu l'utilisation de la représentation condensée utilisant des itemsets aux propriétés particulières, les itemsets δ-libres et leur fermeture [BBR03]. Cette représentation, calculée grâce à l'algorithme AC-MINER [BBR00], permet par la suite de générer un ensemble concis de règles d'association dites δ-fortes.

1.5.3 Découverte de connaissances à partir de règles d'association

Quel que soit l'algorithme utilisé, le nombre de règles qui vont être générées dépend fortement de la densité de la matrice, du nombre d'attributs (colonnes) et du seuil de fréquence minimale. Une fois ces paramètres définis, on collecte un ensemble de règles d'association. Ces règles doivent ensuite être analysées par l'expert, afin qu'il puisse sélectionner les règles qu'il jugera pertinentes.

La principale difficulté de cette phase d'analyse provient, d'une part, du très grand nombre de règles d'association extraites et, d'autre part, de la redondance par rapport au domaine d'application, portée par un grand nombre de règles. Ainsi les règles réellement intéressantes sont littéralement noyées dans la masse des résultats issus de l'algorithme d'extraction. Les contributions apportées à ce problème restent limitées dans leur efficacité, notamment lorsqu'il s'agit de gérer efficacement la redondance intrinsèque au domaine d'application. Une autre voie de recherche consiste à utiliser les connaissances du domaine pour faire ressortir les règles présentant une information qui apparaît comme contradictoire. Les travaux de S. Jaroszewicz et al. [JS04], en particulier, ont posé les premiers jalons quant à la possible utilisation d'un Réseau Bayésien comme modèle des connaissances du domaine. Dans leurs travaux, ce réseau est exploité pour mettre en évidence des ensembles d'attributs potentiellement intéressants au regard des connaissances déjà modélisées par le réseau bayésien.

Dans [FDMB06a], nous sommes partis de cette proposition pour décrire une méthodologie d'extraction de règles d'association pertinentes : sous l'hypothèse qu'un réseau bayésien capture de la connaissance experte sur certaines dépendances entre les variables du domaine, il est alors possible de présenter des règles d'association plus intéressantes. Ceci étant, la disponibilité et la mise à jour de tels modèles peuvent constituer de nouveaux verrous. Pour un expert dans un domaine d'application, par exemple les interruptions opérationnelles des avions, construire mais aussi exploiter et faire évoluer une modélisation par réseau bayésien n'est pas simple. Cette difficulté est aggravée lorsqu'il faut traiter de grands volumes de données issues de sources d'informations souvent hétérogènes et impliquant de très nombreuses variables. Nous avons donc étudié plus précisément [FDMB06b] les interactions entre l'expert d'une part et le réseau bayésien qui modélise une partie de sa connaissance d'autre part. La validation de ces travaux a été poursuivie dans [FDBM06] où nous avons étudié l'application plus systématique de notre approche sur des données simulées.

Chapitre 2

État de l'art sur la découverte de règles d'associations pertinentes

Ce chapitre dresse un état de l'art des approches pour la découverte de connaissances par le biais de l'extraction de règles d'association qui se révèlent pertinentes aux yeux d'un expert. Nous allons présenter plusieurs axes de recherche relatifs à cette problématique. Pour cela il apparaît important de commencer par une présentation des règles d'association ainsi que des problèmes posés par les phases d'extraction et d'analyse des résultats. Pour chaque grande catégorie de problèmes nous détaillons les contributions issues de la littérature. Cela nous mènera, en particulier, à décrire l'exploitation de modèles de connaissance – tels que les Réseaux Bayésiens – dans le cadre de la découverte de règles d'association réellement intéressantes.

2.1 Introduction à la problématique

2.1.1 Philosophie de l'extraction de règles d'association

Les techniques d'extraction de règles d'association ont été introduites pour la première fois en 1993 [AIS93] dans le but de calculer des motifs fréquents à partir de données dites « transactionnelles » [1]. Historiquement, on cherchait à mettre en évidence des comportements valides et inattendus d'acheteurs, à partir de grandes bases de données de transactions. Ces comportements étant présentés sous la forme de règles d'association.

Plus généralement, les règles d'association sont utilisées lorsque l'on veut découvrir des ensembles de couples (attribut, valeur) – appelés *itemsets* – qui apparaissent

[1] En anglais la dénomination fréquemment employée est : *market basket data*.

fréquemment ensemble au sein d'un même jeu de données, et sont liés entre eux par une relation d'association. L'observation des événements de la partie gauche de la règle est souvent associée à l'observation des événements de la partie droite.

Une règle d'association s'exprime de la façon suivante, A → B, où :

- A est la *partie gauche* (aussi appelée antécédent, prémisse ou corps de la règle) ; elle représente les données examinées.
- B est la *partie droite* de la règle (ou conséquent) ; c'est la propriété qui a été découverte en relation avec la partie gauche de la règle,
- A et B sont tous deux appelés des *itemsets*. Ils représentent un ensemble non vide de couples (attribut, valeur) observés sur les données.

L'exemple classique montre la découverte de la règle d'association « couches → bières », à partir d'une base de données constituée des différents paniers achetés dans une grande surface. Cette règle d'association nous dit que lorsqu'un client achète des couches il achète souvent de la bière. Une possible explication de cette découverte serait que les jeunes hommes en charge d'un enfant en bas âge n'ont plus assez de temps pour sortir boire de la bière, ils profiteraient donc de l'achat d'ustensiles pour bébés pour satisfaire leur soif ! Cette découverte décrit un comportement totalement inattendu, mais pourtant valide, dans le sens où cette association s'appuie sur un nombre suffisant d'enregistrements. Elle est de plus potentiellement intéressante, en effet le responsable des ventes peut, par la suite, utiliser cette règle pour optimiser sa stratégie commerciale, par exemple en modifiant la proximité des produits concernés ou leurs prix de vente.

Les règles d'association ont été étudiées et employées dans de nombreux contextes et pour des domaines d'applications variés. On peut citer notamment : la détection d'intrusion ou d'attaques sur un réseau informatique [Kle99, LSM00], la fouille de grands volumes de données textuelles [HC99], l'analyse de données atmosphériques [PKS+03], dans le domaine de la génomique [SSO+97], ou encore pour faciliter la découverte de modules fonctionnels dans des associations de protéines [XHD+05].

2.1.2 Représentation binaire des données

Cette section reprend la terminologie utilisée dans le domaine de l'analyse de règles d'association et présente une définition formelle de la problématique d'extraction de règles.

Définition 2.1 (Base de données binaire) *Soit $\langle T_{id}, Items\rangle$ le schéma d'une base de données binaire. Items est un ensemble de n attributs*[2] *$\{A, B, C, \ldots\}$. L'attribut*

[2]Dans le cadre d'une base de données binaire, et dans notre contexte, on s'intéresse uniquement à la **présence** d'un événement dans les données. Ainsi, pour simplifier la notation, on désignera un item uniquement par son attribut, sa valeur étant supposée à 1.

*T_{id} de type entier est la clef de la relation. Une base de données binaire qui instancie ce schéma est un ensemble de lignes (ou transactions) dont chacune est composée d'un identifiant unique (noté t_i) et d'un sous ensemble de Items, noté t_i.*item.

La figure 2.1 présente deux représentations équivalentes d'une base de données bd de schéma $\langle T_{id}, Items\rangle$. Ici, l'ensemble *Items* est égal à $\{A, B, C, D, E\}$ et la base comporte un total de 6 transactions. Dans le tableau de droite chaque enregistrement est représenté par l'ensemble des attributs observés, dans la partie gauche ces enregistrements sont présentés sous une forme « binaire » (un 1 représente la présence d'un attribut, un 0 son absence). On se référera fréquemment à ces données lors des exemples présentés dans ce chapitre.

T_{id}	t_i.item
1	{A,B,C,D,E}
2	{A,B,C,E}
3	{C}
4	{B,C
5	{A,B,C,D}
6	{A,B,C}

T_{id}	A	B	C	D	E
1	1	1	1	1	1
2	1	1	1	0	1
3	0	0	1	0	0
4	0	1	1	0	0
5	1	1	1	1	0
6	1	1	1	0	0

FIG. 2.1 – Exemple de base de données transactionnelles T (à gauche), et représentation binaire associée (à droite).

2.1.3 Itemsets et règles d'association

Définition 2.2 (Itemset, règle d'association) *L'ensemble $S = \{i_1, i_2, \ldots, i_k\} \subseteq Items$ est appelé itemset, ou k-itemset s'il contient k éléments. Par exemple $\{A, B, C\}$ est un 3-itemset; pour simplifier l'écriture on utilisera également la notation ABC. L'ensemble des itemsets est noté 2^{Items}.*
Une règle d'association $\mathbb{R}$ est un motif $X \rightarrow Y$ où X et Y sont des itemsets sur Items tels que $Y \neq \emptyset$ et $X \cap Y = \emptyset$. X est appelé le corps, l'antécédent ou la partie gauche de la règle et Y la tête, le conséquent ou la partie droite.

Une règle représente une association entre deux itemsets, cette association peut être quantifiée par un ensemble de mesures. Les deux mesures les plus classiques sont la fréquence et la confiance.

Définition 2.3 (Support, fréquence, confiance) *Soit bd une base de données binaire de schéma $\langle T_{id}, Items\rangle$. Soit S un itemset ($S \subseteq Items$), une transaction t supporte S si $S \subseteq t$.*item. *Le support de S dans bd, noté* supp(S, bd)*, est l'ensemble*

des transactions de bd qui supportent S.

$$\text{supp}(S, bd) = \{t \in bd \mid S \subseteq t.\text{item})\}.$$

La fréquence absolue de S dans bd est définie comme le cardinal du support de S :

$$\text{Freq}_a(S, bd) = |\text{supp}(S, bd)|$$

La fréquence relative est la proportion des lignes qui supportent S par rapport à l'ensemble des enregistrements de la base,

$$\text{Freq}_r(S, bd) = \frac{|\text{supp}(S, bd)|}{|bd|}$$

Soit R la règle d'association telle que $X \Rightarrow Y$. Le support et la fréquence de R sont définis comme le support et la fréquence de $X \cup Y$. La confiance de R dans bd est donnée par :

$$\text{conf}(X \rightarrow Y, bd) = \frac{\text{Freq}(X \rightarrow Y, bd)}{\text{Freq}(X, bd)}$$

*La **fréquence** et la **confiance** sont deux mesures permettant d'évaluer la force d'une règle. Une règle d'association est dite* `exacte` *lorsque sa confiance est égale à* 1.

Exemple. Considérons la règle AB → E extraite à partir de la base de données bd (figure 2.1). Comme le cardinal du support de ABE est de 2 et que le nombre total de transactions est de 6, la fréquence absolue de cette règle est de $2/6$ soit $0,33$. On obtient la confiance en divisant la fréquence de ABE par celle de AB. Cela nous donne $2/4$ soit $0,5$.

La mesure de fréquence nous donne une information importante sur les règles. Si la fréquence est très faible cela peut vouloir dire que la conjonction d'événements qu'elle représente est le fruit du hasard. D'un autre côté les règles ayant une fréquence très élevée ont de grandes chances d'être déjà connues par les experts du domaine étudié. Comme on va le voir, la fréquence possède une propriété intéressante que l'on va exploiter pour pouvoir extraire efficacement les règles.

La confiance d'une règle X → Y mesure la fiabilité de l'implication entre X et Y. Plus grande est la confiance et plus grande sera la probabilité que Y apparaissent dans les mêmes transactions que X. Cependant cette mesure est a manipuler avec beaucoup de précautions. En effet, la notion d'association portée par la règle n'est pas synonyme de causalité. Elle suggère simplement une forte co-occurence des éléments de la partie gauche de la règle avec ceux de la partie droite.

Les sections suivantes vont présenter un état de l'art – orienté par notre problématique – tout d'abord sur les techniques d'extraction d'itemsets fréquents et la génération des règles d'association, puis sur les différentes approches pour la sélection de règles réellement pertinentes pour l'utilisateur.

2.2 Exploiter les mesures d'intérêt objectives

2.2.1 Définition du problème

Après avoir établi un cadre formel aux règles d'association, nous allons maintenant nous intéresser à la phase d'extraction de ces règles, ainsi qu'aux différentes problématiques qui en découlent.

Une approche naïve consisterait à calculer le support et la confiance de toutes les règles possibles. Cette approche n'est pas réalisable car le nombre de règles $\mathcal{R}$ pouvant être extraites à partir d'une base de données est exponentiel en fonction du nombre d'items d qui composent le jeu de données.

$$\mathcal{R} = 3^d - 2^{d+1} + 1$$

Ainsi en appliquant cette formule sur un petit jeu de données constitué de 6 items (figure 2.1), on se rend compte qu'il est possible de générer $3^6 - 2^7 + 1$ soit 602 règles différentes. Cependant de nombreuses règles ont une fréquence ou une confiance faible ou nulle. Ces règles ne sont pas intéressantes pour l'utilisateur final ; exploiter ces deux mesures (fréquence et confiance), en fixant par exemple des seuils minimaux pour l'extraction, permettrait d'éviter de nombreux calculs inutiles ainsi que de présenter à l'utilisateur un trop grand nombre de règles.

Le problème de l'extraction peut donc se reformuler de la façon suivante : étant donné un ensemble de transactions, découvrir toutes les règles qui ont une fréquence supérieure à $minfreq$ et une confiance supérieure à $minconf$, où $minfreq$ et $minconf$ correspondent respectivement aux seuils de fréquence et de confiance fixés par l'utilisateur.

Pour résoudre ce problème les approches classiques fonctionnent en deux étapes. Tout d'abord, une phase d'**extraction des *itemsets fréquents*** et de leur support. Les itemsets fréquents sont tous les ensembles de couples (attribut, valeur), qui satisfont un seuil de fréquence minimale $minsup$ spécifié par l'utilisateur. Puis, la deuxième étape, consiste en la **génération des règles d'association** de forte confiance (supérieure à $minconf$) à partir des itemsets fréquents et de leur support extraits à l'étape précédente.

La première difficulté découle du fait que l'extraction des itemsets demande des temps de calculs importants, puisqu'on doit généralement effectuer plusieurs passes sur les données pour arriver à calculer le support des itemsets, il s'agit donc d'être en mesure de calculer ces itemsets quel que soit le jeu de données abordé.

Le deuxième point concerne la génération des règles ainsi que leur évaluation. Cette phase a pour objectif d'éliminer le plus grand nombre possible de règles inin-

téressantes, ou redondantes entre elles, et vis-à-vis du domaine d'application. Il faut donc pouvoir définir un critère d'intérêt des règles. On va commencer par revenir sur la fréquence et la confiance, puis on évoquera différentes mesures d'intérêt qui ont été proposées dans la littérature. Nous qualifierons ces critères d'*objectifs* car ils s'appuient uniquement sur les données pour évaluer la qualité d'une règle ; ils ne prennent donc pas en compte le jugement de l'expert ou les connaissances du domaine.

2.2.2 Le cas particulier de la mesure de fréquence

Dans un premier temps nous allons voir comment la mesure de fréquence est utilisée pour permettre une extraction efficace des itemsets. Pour cela on définit la propriété de fréquence d'un itemset, relativement à un seuil γ.

Définition 2.4 (Itemset γ-fréquent) *Un itemset S est γ-fréquent sur bd s'il satisfait la contrainte de fréquence minimale γ. L'ensemble des itemsets γ-fréquents sur bd est donné par*

$$\text{Freq}(bd, \gamma) = \{X \subseteq Items | \text{Freq}(X, bd) \geq \gamma\}$$

Cette définition s'étend naturellement aux règles d'association.

La propriété de fréquence d'un itemset, qui est à la base de l'efficacité des algorithmes d'extraction, s'exprime de la façon suivante :

Proposition 2.1 *La fréquence est une fonction décroissante par rapport à l'inclusion ensembliste. Soit bd une base de données binaire, S et T deux itemsets, alors*

$$S \subseteq T \Rightarrow \text{Freq}(S, bd) \geq \text{Freq}(T, bd)$$

De ce fait, pour un itemset S donné et un seuil de fréquence γ :

- Si S est γ-fréquent alors tout sous-ensemble de S est aussi γ-fréquent.
- Inversement, si S n'est pas γ-fréquent, alors tout sur-ensemble de S ne sera pas γ-fréquent.

Plus généralement on définit les notions de contraintes monotones et anti-monotone de la façon suivante :

Définition 2.5 (Contrainte monotone, anti-monotone) *Une contrainte C_a est anti-monotone si et seulement si pour chaque itemset S, si S ne satisfait pas C_a, alors aucun de ses sur-ensembles ne satisfait C_a.*

Réciproquement C_m est monotone si et seulement si pour chaque S, si S satisfait C_m, alors chacun de ses sur-ensembles satisfait C_m.

La contrainte de *fréquence minimale* est donc une contrainte anti-monotone.

Pour comprendre, plus concrètement, l'impact de cette propriété sur l'extraction d'itemsets fréquents, prenons la base de données transactionnelles bd introduite précédemment, et l'ensemble $Items = \{A, B, C, D, E\}$ associé. Pour notre exemple de génération des itemsets fréquents nous prendrons un seuil de fréquence absolue minimale égal à 2.

Items	Support	Fréquence
{B}	$\{t_1, t_2, t_4, t_5, t_6\}$	5
{A}	$\{t_1, t_2, t_5, t_6\}$	4
{A,B,C}	$\{t_1, t_2, t_5, t_6\}$	4
{A,B,E}	$\{t_1, t_2\}$	2
{B,D}	$\{t_1, t_5\}$	2
{B,C,D}	$\{t_1, t_5\}$	2

TAB. 2.1 – Itemsets fréquents ($minfreq = 2$) extraits à partir de la base bd.

Règles d'association	Fréquence	Confiance
B	0	5
A	0	4
A,B,C	0	4
A,B,E	0	2
B,D	0	2
B,C,D	0	2

TAB. 2.2 – Exemples de règles d'association générées à partir des itemsets fréquents extraits (tableau 2.1).

La figure 2.2 présente le treillis des itemsets obtenu à partir de la base de données bd (figure 2.1). Cette représentation permet de visualiser l'impact de l'application de la propriété d'anti monotonie pour la contrainte de fréquence minimale. Par exemple l'itemset `DE` n'est pas fréquent, ce qui nous permet d'éliminer tous les itemsets de niveau supérieur qui contiennent `DE` : `CDE`, `BDE`, `ADE`, `ABDE`, `ACDE`, `BCDE` et `ABCDE`.

Les tableaux 2.1, et 2.2 nous montrent l'ensemble des itemsets fréquents calculés pour une fréquence absolue supérieure ou égale à 2, ainsi que quelques exemples de règles d'associations pouvant être générées à partir de ces itemsets.

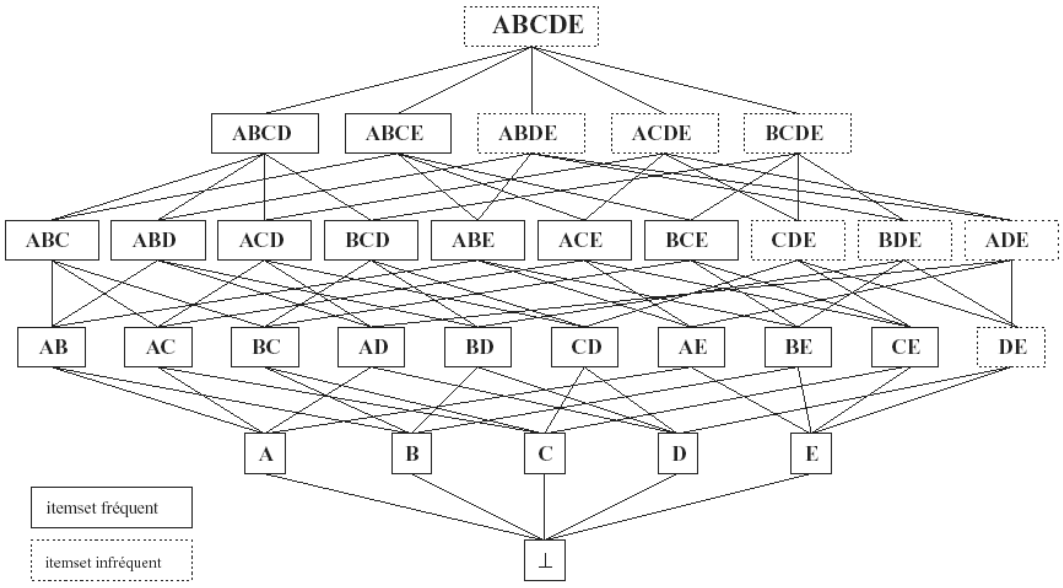

FIG. 2.2 – Treillis des *itemsets* et partition des itemsets fréquents.

2.2.3 Algorithmes d'extraction de tous les itemsets fréquents

APRIORI [AMS+96] est le premier algorithme efficace pour l'extraction d'itemsets fréquents, il procède en deux temps :

1. On recherche les itemsets fréquents, ceux dont la fréquence est supérieure au seuil $minfreq$, en effectuant un parcours en largeur du treillis des et en calculant les fréquences par comptage dans la base, ce qui impose une passe sur la base à chaque niveau du treillis ;
2. Puis, pour chaque itemset fréquent X, on conserve les seules règles du type $X \setminus Y \rightarrow Y$, avec $Y \supset X$, dont la confiance dépasse le seuil $minconf$. On remarque que les règles ainsi conservées ont nécessairement une valeur de confiance supérieure au seuil de fréquence, dans la mesure où $\mathcal{F}(\mathrm{A} \rightarrow \mathrm{B}) > \mathrm{conf}(\mathrm{A} \rightarrow \mathrm{B})$

Ainsi, à chaque niveau du treillis cet algorithme exploite l'anti-monotonicité de la contrainte de fréquence minimale pour ne traiter qu'une partie de l'ensemble des itemsets candidats. La fonction de génération des candidats détermine quelle partie du treillis va être explorée en éliminant les candidats non-fréquents, ainsi que tout leurs sur ensembles.

Si ce type d'approche est efficace pour traiter des données faiblement corrélées (comme les données transactionnelles pour lesquelles il était destiné), les performances chutent terriblement sur des données plus denses et corrélées [BMUT97].

En effet, le principe même de cet algorithme est d'extraire *tous* les itemsets qui satisfont le seuil de fréquence spécifié par l'utilisateur. Ainsi, toute la difficulté de l'extraction des itemsets fréquents consiste à identifier le plus efficacement possible la bordure entre les itemsets fréquents et les itemsets non-fréquents dans le treillis des itemsets [HGN00].

Ainsi, on va définir la frontière positive, respectivement négative de l'ensemble des itemsets fréquents.

Définition 2.6 (Frontière positive, négative) *Soit $\mathcal{S}$ une collection d'itemsets. La frontière) positive de $\mathcal{S}$ est la collection des itemsets les plus spécifiques de $\mathcal{S}$,*

$$Bd^{+}(\mathcal{S}) = max_{\preceq}(\mathcal{S}) = \{\varphi \in \mathcal{S} | \nexists\ \mu \in \mathcal{S}, \varphi \prec \mu\}$$

Où $\preceq$ est une relation d'ordre partiel sur les itemsets telle que l'itemset φ est plus général que μ si $\varphi \preceq \mu$.
La frontière négative de $\mathcal{S}$ est ma collection des motifs les plus généraux qui ne sont pas dans $\mathcal{S}$,

$$Bd^{-}(\mathcal{S}) = min_{\preceq}(2^{Items} \setminus \mathcal{S}) = \{\varphi \notin \mathcal{S} | \nexists\ \mu \notin \mathcal{S}, \mu \prec \varphi\}$$

Apriori est l'algorithme fondateur de cette catégorie, d'autres contributions ont suivi par la suite. En effet, le parcours du treillis peut se faire en largeur ou en profondeur. Dans les deux cas, on peut procéder par comptage direct de la fréquence de chaque itemset dans la base, ou procéder par intersection des deux itemsets qui constituent l'itemset candidat. Différentes améliorations ont été proposées dans la littérature afin d'accélérer l'étape de construction des ensembles fréquents dans certaines situations. Nous présentons, de manière chronologique les plus significatives d'entre elles.

Première amélioration : l'algorithme *AprioriTID* [AS94], vise à limiter les accès directs à la base de données – en pratique les temps d'accès s'avèrent d'autant plus pénalisant que l'algorithme classique effectue une passe sur la base pour chaque niveau du treillis. Pour cela, l'intégralité de la base est mise en mémoire, et à chaque niveau, on représente les transactions par les k-itemsets candidats qu'elle contient. Ainsi, une seule passe est désormais nécessaire. Cependant la sensibilité de cette approche aux jeux de données fortement corrélés restent la même que pour Apriori, et, autre inconvénient, toute la base doit pouvoir tenir en mémoire.

Avec l'algorithme *Partition* [SON95] les auteurs ont investigué l'utilisation des tid-listes (ensemble des tid – ou identifiant unique d'une transaction – associés aux transactions qui contiennent un itemset donné) intermédiaires associées au treillis de chaque partie tienne en mémoire ; dans une première passe, on travaille en largeur et on extrait les tid-listes des itemsets du niveau k pour construire les itemsets fréquents

de chaque partie, par intersection des tid-listes du niveau $k-1$; dans une seconde passe, on vérifie pour chaque ensemble localement fréquent qu'il est bien globalement fréquent.

L'algorithme *Sampling* [Toi96] construit quant à lui l'ensemble des itemsets fréquents ainsi que sa bordure négative, à partir d'un échantillon représentatif de la base. Cette méthode a permet de limiter le risque de non exhaustivité.

Les auteurs de [BMUT97] proposent avec l'algorithme *Dynamic Itemset Counting* un parcours niveau par niveau modifié. Au niveau k dès qu'un itemset a atteint le seuil de fréquence, on introduit les itemsets candidats de niveau $k+1$ qu'il contribue à générer, ce qui diminue le nombre de passes nécessaires sur la base.

Autre approche pour l'extraction des itemsets fréquents : *Eclat* [ZPOL97] effectue cette fois un parcours du treillis en profondeur, par intersection rapide des tid-listes. La procédure étant interrompue dès que l'on est sûr que l'itemset candidat ne peut plus être fréquent.

L'algorithme FP-Growth [HK00] améliore quant à lui les capacités d'extraction des itemsets fréquents, ainsi que les performances globales, en associant une structure spécifique des données de transaction –intitulée *FP-tree*– avec une recherche en profondeur dans le treillis.

En parallèle a ces différentes propositions, un autre axe de recherche a été étudié. Il s'agit de l'extraction d'une partie génératrice des itemsets fréquents et de leur support : les représentations condensées. Ces recherches ont abouti à différents algorithmes on citera notamment *Close* [PBTL99b] et son dérivé *A-Close* [PBTL99a], puis l'algorithme *Pascal* fondé sur le comptage par inférence des ensemble clés [BTP+00]. Les ensembles fermés et les ensembles libres ont aussi été étudiés par J.-F. Boulicaut et al. [BB00, BB00] qui ont étendu ces notions à celles d'ensembles δ-fermés et δ-libres.

Les approches visant à calculer des représentations condensées ont l'avantage de réduire les temps d'extraction – et par là même de rendre la tâche d'extraction possible sur des jeux de données qui ne pouvaient jusque là pas être abordés par des approches de type Apriori. On reviendra plus en détails sur ces contributions dans la section 2.3.

2.2.4 L'approche support-confiance

L'approche intitulée « support-confiance » fait intervenir les mesures de fréquence et de confiance pour évaluer l'intérêt potentiel d'une règle d'association. Dans cette optique, l'utilisation d'un seuil minimal de fréquence a pour intérêt de rendre praticable l'algorithme d'extraction, du fait de la propriété d'antimonotonie du treillis des itemsets fréquents. La confiance doit permettre, quant à elle, de sélectionner les règles potentiellement intéressantes parmi celles qui satisfont à la condition de fréquence.

La condition de fréquence qui est le moteur même du processus d'extraction écarte les règles ayant une faible fréquence alors que certaines peuvent avoir une très forte confiance et présenter un réel intérêt. Comme nous l'avons précisé, les algorithmes classiques ne permettent pas l'extraction à des seuils de fréquence intéressants pour l'utilisateur.

La mesure de confiance est une fonction d'évaluation objective classique. Cependant, comme le montre l'exemple ci-dessous, elle ne permet pas, à elle seule, de garantir la qualité d'une règle d'association. Elle peut de plus être la source d'erreurs d'interprétation des résultats. Prenons le cas où la confiance d'une règle A $\rightarrow$ B est égale à la probabilité de B. Selon la définition de la confiance on a alors $P(B/A) = P(B)$, c'est à dire une indépendance entre A et B. Cette règle qui peut avoir une forte confiance ne nous apporte aucune information, elle ne doit donc pas être jugée comme intéressante.

L'exemple suivant permet d'illustrer ce propos.

Exemple. Supposons que l'on souhaite analyser la relation qui existe entre des personnes achetant les produits A et B. Le tableau 2.3 montre la répartition des achats sur la population étudiée.

	A	$\bar{A}$	$\sum$
B	150	50	200
$\bar{B}$	650	150	800
$\sum$	800	200	1000

TAB. 2.3 – Répartition des achats sur un groupe de 1000 personnes.

À partir de ces données il est possible d'évaluer la règle A $\rightarrow$ B($\mathcal{F} = 0,15; \text{conf} = 0,75$). Les valeurs raisonnablement élevées pour les mesures de fréquence et de confiance nous invitent à penser que les personnes qui achètent le produit A achètent aussi le produit B. Cependant, la part des personnes achetant B, *indépendemment* du fait qu'elles achètent aussi A est de $0,80$ alors que la règle nous dit que la proportion de personnes consommant A et B est inférieure, puisqu'elle est égale à $0,75$. La règle A $\rightarrow$ B qui semblait intéressante s'avère donc trompeuse.

L'utilisation d'autres mesures paraît donc nécessaire pour repérer les règles réellement intéressantes. Dans la section suivante nous évoquons différentes contributions issues de la littérature sur les mesures d'intérêt dites « objectives ».

2.2.5 Autres mesures de l'intérêt objectif d'une règle d'association

Nous avons vu que les algorithmes du type APRIORI, fondés sur la fréquence et la confiance des règles, ont apporté une première solution au problème de l'extraction

de règles, mais ils produisent une trop grande masse de règles, sélectionnant certaines règles sans intérêt et ignorant des règles intéressantes.

Ainsi, de nombreux travaux de recherche visent à définir de nouvelles mesures pour compléter le support et la confiance.

Le sujet est trop vaste pour que toutes ces mesures soient énumérées ici. On peut néanmoins citer les mesures les plus utilisées. Le *Lift*, par exemple, est une façon de régler le problème d'interprétation de la confiance que nous avons vu dans l'exemple précédent. Il se définit de la façon suivante :

$$Lift = \frac{\text{conf}(\text{A} \rightarrow \text{B})}{\mathcal{F}(B)} = \frac{\mathcal{F}(AB)}{\mathcal{F}(A) \times \mathcal{F}(B)}$$

Le fait de faire intervenir la fréquence de B, par rapport à la confiance, revient à comparer la fréquence de l'itemset AB constatée par rapport à sa fréquence sous l'hypothèse d'indépendance statistique entre A et B. Ainsi, en appliquant le *Lift* à notre exemple (tableau 2.3), on obtient $Lift(\text{A} \rightarrow \text{B}) = \frac{0,15}{0,2 \times 0,8} = 0,9375$. Ce résultat peut s'interpréter comme une faible corrélation négative entre les acheteurs des produits A et B.

Le coefficient de corrélation de Pearson, la J-Mesure , ou encore l'intensité d'implication ([BKGG04]) sont d'autres exemples de mesures objectives pouvant être calculées sur les règles d'association. Chacune d'entre elles permet d'évaluer un critère statistique bien précis qui pourra avoir un intérêt ou non, aux yeux de l'expert.

Le lecteur désirant une description plus détaillée de ces mesures pourra se référer aux travaux de synthèses réalisés sur le sujet, en particulier la thèse de [Azé03], ou encore les travaux de [LT04, GCB+04].

De plus, face au grand nombre de mesures proposées et à la multitude de règles candidates qu'il faut analyser, un autre problème émerge, celui du choix des mesures d'intérêts les plus adaptées a un cas d'application donné. Autrement dit, il devient important de définir des critères permettant d'évaluer ces mesures. Ainsi [LMV+04], présente une approche multicritères pour l'aide à la décision dans le choix des mesures à utiliser. [LPT04] propose une méthode de validation qui utilise les outils de la théorie de l'apprentissage statistique, notamment la VC-dimension. L'objectif de cette dernière contribution est de permettre la construction de bornes uniformes non asymptotiques pour toutes les règles et toutes les mesures simultanément.

2.2.6 Conclusion

La génération d'itemsets par le biais d'un parcours niveau par niveau exploitant *uniquement* la contrainte de fréquence minimale, n'est *pas* efficace sur des jeux de don-

nées de réels. L'approche de type APRIORI a néanmoins été populaire car les critères utilisés – fréquence et confiance – *semblent* faciles à mettre en place et à interpréter. L'algorithme en lui même demeure intéressant car il est l'implémentation d'un algorithme *générique* de parcours niveau par niveau (pour la contrainte de fréquence), simple a à mettre en place et a adapter pour d'autres contraintes anti-monotones.

Nous avons aussi vu qu'il était nécessaire de donner à l'utilisateur les moyens de choisir d'autres critères prenant en compte la nature particulière des règles d'association tout en étant les plus adaptés au problème traité. La visualisation et le classement des règles sur la base de différents critères est une piste intéressante. Cependant l'approche classique connaît plusieurs limitations qui nous abordons dans les sections suivantes :

- Les algorithmes qui suivent cette approche ne sont pas efficaces sur des données fortement corrélées et/ou à des seuils de fréquence intéressants pour l'utilisateur.
- Les règles obtenues sont très nombreuses. La découverte de règles réellement intéressantes est donc d'autant plus délicate que de nombreuses règles s'avèrent inintéressantes et/ou redondantes.
- Les mesures dites objectives ne permettent pas, par définition, de s'affranchir de la redondance liée au domaine d'application.

2.3 Éliminer la redondance des règles fréquentes et valides

2.3.1 Définition du problème

La génération des règles d'association par le biais d'un algorithme de type APRIORI génère un grand nombre de règles. Une bonne partie de ces règles sont redondantes. Par exemple, à partir de la base de donnée T il est possible de générer les règles suivantes, de fréquence égale à 2 et de confiance 1 : E $\rightarrow$ C, E $\rightarrow$ BC, AE $\rightarrow$ C et AE $\rightarrow$ BC. Ces règles sont toutes valides mais on voit bien qu'il est possible de les regrouper en une seule et unique règle exprimant la même information, la règle : AE $\rightarrow$ BC.

De manière plus générale, on définit la redondance de la façon suivante :

Définition 2.7 (Redondance d'une règle d'association) *Une règle d'association $R : X \rightarrow Y$ est redondante s'il existe une autre règle $R' : X' \rightarrow Y'$, telle que $X \subseteq X'$, $Y \subseteq Y'$ et, le support et la confiance de R et R' sont identiques.*

A l'échelle d'une application réelle un nombre important de règles générées sont redondantes et viennent parasiter la phase d'exploration des résultats, rendant d'autant

plus difficile la découverte de règles intéressantes.

Cette redondance que nous qualifions d'*intrinsèque* peut être abordée *a posteriori*, par un post-traitement sur l'ensemble des règles obtenues ; à ce sujet on pourra consulter les contributions suivantes [LHM99, JS02]. Une autre façon d'envisager le problème consiste à générer les règles de telles sortes qu'elles ne soient pas redondantes entre elles [Pas00]. Nous allons nous intéresser plus particulièrement à ce dernier type d'approche en introduisant les représentations condensées.

2.3.2 Représentations condensées des itemsets fréquents

Le concept des *représentations condensées* a été introduit en 1996 [MT96], dans le cadre plus général de la fouille de données. On a évoqué dans la section précédente l'intérêt que pouvait présenter ces types de représentations pour le calcul des itemsets fréquents et des règles d'association :

- Elles peuvent être extraites plus efficacement que l'ensemble des itemsets fréquents, tout en permettant de les régénérer (pas de perte d'information).
- Elles contiennent la même information que les collections d'itemsets fréquents, tout en étant plus concises.

Il existe plusieurs types de représentations condensées [ZO98, PBTL98, BB00, BBR00]. Ici, nous détaillons les représentations utilisant les itemsets maximaux fréquents, puis, plus particulièrement, les représentations utilisant les fermés, les libres et les δ-libres , ainsi que les différents algorithmes qui permettent de les générer.

Extraction d'itemsets maximaux fréquents

Pour obtenir tous les itemsets qui satisfont une contrainte anti-monotone (comme la contrainte de fréquence minimale) il nous suffit de calculer la frontière positive de cette collection. Les itemsets ainsi trouvés sont appelés *itemsets maximaux*. En effet, si un itemset X vérifie une contrainte anti-monotone, alors tout itemset plus général que X vérifiera également cette contrainte.

Définition 2.8 (Itemset maximal fréquent) *Un itemset maximal fréquent est un itemset fréquent dont aucun de ses sur-ensembles immédiats n'est fréquent.*

Exemple. Considérons le treillis des itemsets présenté dans la figure 2.2. Pour chaque itemset fréquent situé sur la bordure positive, on regarde si tous ses sur-ensembles immédiats sont infréquents. Au final les seuls itemsets maximaux fréquents de notre exemple sont : $ABCD$ et $ABCE$.

Ce type de représentation regroupe donc les techniques qui cherchent à calculer directement cette frontière positive, vis-à-vis d'une contrainte anti-monotone donnée.

Max-clique et Max-eclat [ZPOL97], Max-miner [Bay98], Pincer-search [LK98], Depth-Project [AAP00] sont autant d'approches différentes cherchant à arriver le plus vite possible à la frontière positive, en n'explorant pas de manière exhaustive les différents niveaux du treillis.

Leur intérêt devient évident pour des applications où il est nécessaire d'extraire de très longs itemsets. En effet si dans une application spécifique il existe un itemset fréquent de taille n, un algorithme de parcours niveau par niveau devra d'abord considérer les 2^n sous-ensembles de cet itemset avant de pouvoir le trouver. En pratique, on considère que ce n'est plus praticable quand n est supérieur à 20.

Cependant ces algorithmes n'ayant pas le même objectif qu'APRIORI, ils ne permettent pas de connaître la fréquence de tous les itemsets fréquents. De ce fait, il est impossible de générer toutes les règles d'association car on ne connaît pas précisément les fréquences de tous les sous itemsets. Il est évidemment possible de partir des itemsets maximaux pour dériver tous les sous-ensembles et leur fréquence, mais cette stratégie n'est pas efficace : il a été montré [BTP+00] que ce calcul était au moins aussi coûteux que d'utiliser l'algorithme APRIORI.

Extraction d'itemsets fermés fréquents, libres fréquents

Définition 2.9 (fermeture, fermé) *La fermeture d'un itemset fermé S est le plus grand sur ensemble de S qui a la même fréquence que S. Un itemset S est fermé, ou clos, s'il est égal à sa propre fermeture.*

Il découle de cette définition qu'un itemset fermé est un itemset dont la fréquence est différente de tous ses sur ensembles. De la même façon on définit ce qu'est un itemset libre :

Définition 2.10 (libre) *Un itemset est libre si il n'est pas inclus dans la fermeture d'un de ces sous ensembles stricts.*

On introduit maintenant le concept de classe d'équivalence pour la fermeture :

Définition 2.11 (classe d'équivalence [BTP+02]) *Deux itemsets S et T sont équivalents dans la base de données bd s'ils ont même la fermeture dans bd.*

Pour mieux appréhender la notion de classe d'équivalence, d'itemsets fermés et d'itemsets libres on peut se reporter à la figure 2.3. Le treillis représenté dans cette

figure a été instancié à partir des données présentées dans le tableau 2.1. Les itemsets en gras représentent les itemsets fermés, ils sont les itemsets maximaux des classes d'équivalence. Les itemsets en italique sont les itemsets libres, ils sont les itemsets minimaux des classes d'équivalence. Les différentes classes d'équivalences relatives à la fermeture sont entourées.

Une première lecture de ce treillis permet de se rendre compte qu'il nous suffit de connaître l'ensemble des itemsets libres et leurs fermetures pour qu'il soit possible de générer l'ensemble des itemsets fréquents. C'est pourquoi on parle de représentations condensées : il n'y a pas de perte d'information sur les itemsets et leur fréquence, mais il y a une possible réduction de nombre de motifs à prendre en compte.

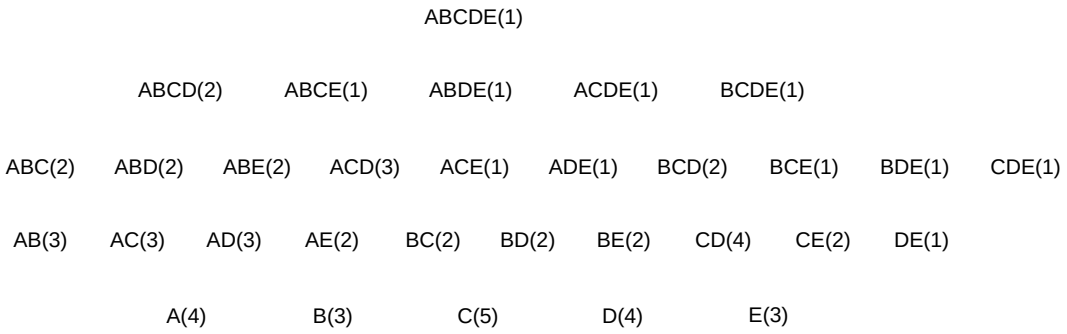

FIG. 2.3 – Treillis des itemsets

Pour réaliser l'extraction des itemsets fermés fréquents différents algorithmes ont été proposés. En particulier, on va différencier deux types d'approches : l'approche niveau par niveau [PBTL99a, BTP+02] et l'approche en profondeur d'abord [PHM00, ZH02].

Approche niveau par niveau. La notion de classe d'équivalence introduite précédemment nous permet de comprendre facilement les points suivants :

- Tous les sous ensembles d'un itemset libre sont libres. La propriété de « liberté »associée à un itemset est dite anti-monotone.
- Par contre, la propriété de « fermeture » n'est ni monotone, ni anti-monotone, on peut d'ailleurs s'en rendre compte en examinant la figure 2.3.

Dans ce cas, comment extraire efficacement les itemsets fermés ? Heureusement, il a été montré que pour extraire les itemsets fermés fréquents il suffisait d'extraire les itemsets libres fréquents et de calculer leur fermeture.

Les algorithmes CLOSE [PBTL98] et PASCAL [BTP+02] sont deux exemples de ce type d'approche utilisant cette propriété. Le parcours du treillis est effectué niveau par niveau, mais au lieu de retourner les itemsets fréquents, on retourne les itemsets libres

fréquents. Une fois les itemsets libres fréquents récupérés, on calcule leurs fermetures et on obtient la collection des fermés fréquents. Ainsi, par rapport à APRIORI, tous les itemsets non libres sont élagués, il y a donc une diminution du nombre d'itemsets traités. Dans le pire des cas cependant, si tous les itemsets sont libres, on va parcourir autant d'itemsets qu'avec APRIORI, mais avec un léger sur-coût engendré par la vérification de la contrainte sur les itemsets libres. En pratique, sur des données fortement corrélées, beaucoup d'itemsets sont libres et l'élagage se révèle efficace.

Approche en profondeur d'abord. Cette classe d'algorithme effectue un parcours en profondeur de l'espace de recherche. Lorsque l'algorithme a fini de traiter un itemset S il examine ses fils par ordre croissant de fréquence. Les fils de S sont les itemsets de la forme $S \cup i$ ou i est un item qui n'a pas encore été examiné dans d'autres branches de l'arbre constitué par les itemsets.

Closet [PHM00] et *Charm* [ZH02] sont les deux algorithmes les plus efficaces concernant l'extraction des itemsets fermés fréquents par un parcours en profondeur d'abord.

2.3.3 Différentes représentations condensées des itemsets fréquents

Les représentations condensées utilisant les itemsets fermés fréquents que nous introduisons ici ont été présentées pour la première fois par [BB00]. Lorsque les données sont fortement corrélées le gain de taille par rapport à la représentation de tous les itemsets fréquents peut être de plusieurs ordres de grandeur. Comme on le verra, cette collection est aussi très intéressante car elle permet de générer un ensemble concis de règles d'association (section 2.3.4). Une propriété essentielle de ce type de collection a été proposée dans [PBTL99a]. En substance elle nous dit qu'à partir de l'ensemble des itemsets fermés γ-fréquents il est possible de calculer la fréquence de n'importe quel itemset γ-fréquent.

Un autre type de représentation condensée, cette fois utilisant les itemsets libres fréquents, a été présentée dans [BBR00]. Les auteurs montrent que pour retrouver la fréquence de n'importe quel itemset S γ-fréquent, il faut d'abord pouvoir montrer que cet itemset est bien γ-fréquent, c'est-à-dire qu'il n'appartient pas à l'ensemble des itemsets des non γ-fréquent ou qu'il n'est pas libre.

[BBR00] a aussi proposé un autre type de représentation condensée, basée cette fois sur une généralisation des itemsets libres, les itemsets δ-libres.

Définition 2.12 (δ-fermeture, [BBR00]) *Soit δ un entier positif. La δ-fermeture d'un itemset S (notée* $\mathrm{ferm}_\delta(S)$ *est définie par :*

$$\mathrm{ferm}_\delta(S) = \{I \in Items \mid \mathrm{Freq}_a(S) - \mathrm{Freq}_a(S \cup \{I\}) \leq \delta\}$$

Freq_a *est la fréquence absolue.*

En pratique cela signifie qu'on peut se passer de calculer la fréquence de certains itemsets en les approchant par la fréquence de leur δ-fermeture.

De la même façon que pour la propriété de liberté un itemset δ-libre se définit comme suit :

Définition 2.13 (δ-libre) *Un itemset est δ-libre s'il n'est pas inclus dans la δ-fermeture de l'un de ses sous ensembles stricts.*

La contrainte δ-libre est anti-monotone. Elle peut donc être exploitée par un algorithme de type CLOSE. C'est ce que fait l'algorithme MIN-EX[3] qui sera employé tout au long des travaux de thèse pour générer les règles d'association dites δ-fortes.

Définition 2.14 (règle δ-forte) *Soit δ un entier positif, une règle d'association est dite δ-forte si elle admet au plus δ exceptions.*

Pour une règle d'association $X \rightarrow Y$ on qualifie d'exception toute transaction de la base de données qui supporte X mais pas Y.

Tout comme CLOSE, MIN-EX fait un parcours niveau par niveau du treillis des itemsets, mais au lieu de calculer la fermeture il calcule la δ-fermeture, lors de la passe sur les candidats. Si δ vaut 0 MIN-EX fonctionne de la même façon que CLOSE.

2.3.4 Génération d'une collection non redondante de règles d'association fréquentes et valides

Il est admis que la collection de règles générées à partir de tous les itemsets fréquents comporte une large part de redondance. Plutôt que de chercher à supprimer cette redondance *a posteriori*, M.J. Zaki [Zak00], ainsi R. Taouil [TPBL00] et N. Pasquier [PBTL99a, PTB+05], ont montré qu'il était possible de générer une représentation non redondante des règles d'association à partir des itemsets fermés fréquents et des itemsets libres (aussi appelés générateurs).

Dans le cas des travaux de N. Pasquier, cette représentation contient un ensemble de règles d'association ayant une partie gauche minimale et une partie droite maximale (au sens de l'inclusion). Ces règles sont appelées *règles d'association min-max*. Les auteurs pensent que ce type de représentation est le plus pertinent car l'information présentée par les règles est la plus générale possible. Les définitions qui suivent sont inspirées de [PTB+05].

[3] L'implémentation de MIN-EX utilisée est le programme AC-like développé par Jérémy Besson, il est disponible à l'adresse suivante : `http://liris.cnrs.fr/jeremy.besson/AC-like/AC-like.html`

Définition 2.15 (Règle d'association générale) *Une règle d'association $R : X \rightarrow Y$ est plus générale qu'une règle $R' : X' \rightarrow Y'$ si les conditions suivantes sont réunies :*
- *(1) $\mathcal{F}(R) = \mathcal{F}(R')$ et* $\text{conf}(R) = \text{conf}(R')$
- *(2) $X \subset X'$ et $Y \supset Y'$*

R est alors une sur-règle de R' et R' une sous-règle de R

Afin d'illustrer ce propos, examinons les règles présentées dans le tableau 2.4. Elles ont été extraites à partir de la base de données bd présentée précédemment (tableau 2.1).

Num	Règles	Fréquence	Confiance
1	A → BC	4	1,00
2	AE → BC	2	1,00
3	E → B	2	1,00
4	E → C	2	1,00
5	E → BC	2	1,00

TAB. 2.4 – Exemple d'un ensemble de règles d'association pouvant être simplifié

Toutes ces règles sont dites *exactes* car elles ont une confiance de 1. Comme on peut le voir les règles (3) et (4) peuvent être directement déduites des règles (2) et (5). La règle (5) est appelée *règle min-max exacte* selon l'expression utilisée dans [PTB+05]. Il s'agit de la règle la plus générale par rapport à (2), (3) et (4) ; on va donc la conserver.

Définition 2.16 (Règle d'association min-max) *La règle $R : X \rightarrow Y$ est une règle d'association min-max ssi il n'existe pas de règle $R' : X' \rightarrow Y'$ plus générale que R.*

Pour générer l'ensemble des règles min-max, on fait la distinction entre les règles min-max exactes et les règles min-max approximatives (de confiance < 1).

Définition 2.17 (Base min-max exactes) *Soit $\texttt{Ferm}$ l'ensemble des itemsets fermés fréquents, et $\texttt{Libre}_X$ l'ensemble des libres de la même classe d'équivalence que X. Alors, la base[4] des règles min-max exactes est exactement :*

$$MinMaxExact = \{R : Y \rightarrow X \backslash Y | X \in \texttt{Ferm} \wedge Y \in \texttt{Libre}_X \wedge Y \neq X\}.$$

On a vu précédemment un exemple d'extraction d'itemsets libres et de leur fermeture sur les données de bd. En repartant de cet exemple on va générer les règles min-max exactes. Le résultat est présenté dans le tableau 2.5.

[4]Ensemble minimal de règles à partir duquel on peut générer l'ensemble des règles d'association.

Num	Libres	Fermeture	Règle min-max exacte	Fréquence
1	A	ABC	A → BC	4
2	B	BC	B → C	5
3	D	ABCD	D → ABC	2
4	E	ABCE	E → ABC	2

TAB. 2.5 – Ensemble des règles min-max exactes obtenues à partir de la base de données bd.

De même on définit la base des règles min-max approximatives, composée de règles d'association de confiance strictement inférieure à 1.

Définition 2.18 (Base min-max approximative)

$$MinMaxApprox = \{R : Y \rightarrow X \backslash Y | X \in \texttt{Ferm} \wedge Y \in \texttt{Libres} \wedge \text{ferm}(Y) \supset \text{ferm}(X)\}$$

Cela revient à dire que Y est un libre et X un fermé appartenant à une classe d'équivalence différente de celle de Y et telle que $\text{ferm}(Y) \supset X$.

A partir de cette base il est encore possible d'effectuer une réduction sans perte d'information (i.e., en conservant la base). Le but de cette réduction est de sélectionner parmi les règles transitives, celles qui ont la plus grande valeur de confiance (i.e., les règles les plus précises). Une règle min-max approximative de la forme $R : Y \rightarrow X \backslash Y$ est dite *transitive* si il existe un itemset fermé fréquent X' tel que $\text{ferm}(Y) \supset X' \supset X$.

Le tableau 2.6 nous montre un exemple de règles min-max approximatives obtenues à partir des données de bd.

Num	Libres	Fermeture du surensemble	Règle min-max approx	Confiance
1	A	ABC	A → BC	1,00
2	D	ABCDE	D → ABCE	0,50
3	E	ABCDE	E → ABCD	0,50

TAB. 2.6 – Ensemble des règles min-max approximatives générées à partir de bd.

Définition 2.19 (Base min-max approximative non-transitive)

$$MinMaxReduc = \{R : Y \rightarrow X \backslash Y | X \in \texttt{Ferm} \wedge Y \in \texttt{Libres} \wedge \text{ferm}(Y) \lessdot \text{ferm}(X)\}$$

$X \lessdot Y$ désigne le fait que X est un prédécesseur immédiat de Y, i.e. il n'existe pas d'itemset Z tel que $X \supset Z \supset Y$.

2.3.5 Conclusion

Comme on le voit, la question de la génération de règles non redondantes a été bien étudiée, notamment dans le cas de la fermeture d'itemsets. Il peut cependant être intéressant d'étudier ce qui se passe si l'on utilise l'opérateur de δ-fermeture. Pour δ égal à 0 on sait déjà que la δ-fermeture se comporte comme la fermeture. Intuitivement, il paraît alors possible de généraliser la génération de règles min-max non redondantes, lorsque le paramètre δ devient strictement supérieur à 0 (i.e., considérer l'ensemble des itemsets δ-libres et leur δ-fermeture comme une représentation condensée des itemsets fréquents). Une des principales difficultés à prendre en compte est que la *delta*-fermeture, contrairement à la fermeture, n'est pas idempotente (i.e., $\delta - \text{ferm}(X) \neq \delta - \text{ferm}(\delta - ferm(X))$).

Le chapitre 3 sera l'occasion de revenir plus en détails sur les propriétés des itemsets δ-libres et de la δ-fermeture et des règles générées. C'est un problème qui, à ce jour, n'a pas été étudié. Notamment, cela peut être intéressant car le paramètre δ apporte un bon compromis entre rapidité d'exécution, taille et précision des résultats.

A présent que l'on dispose de techniques permettant à la fois de générer un ensemble concis de règles d'association (élimination de la redondance intrinsèque), mais aussi de travailler à des seuils de fréquence bas, sur des données complexes, une autre piste de réflexion se dessine. En effet, pour des applications réelles, les règles présentées sont toujours trop nombreuses et difficiles à analyser. Elles comportent des informations déjà parfaitement connues de l'expert, ou au contraire elles présentent des motifs qui n'ont pas de rapport avec les objectifs de recherche fixés par l'utilisateur. Dans la prochaine section on va donc réfléchir aux techniques permettant d'introduire la subjectivité dans le processus de découverte de règles intéressantes, ainsi qu'à celles dont le but est de limiter la *redondance au domaine d'application*.

2.4 Exploiter la subjectivité pour sélectionner les règles pertinentes

2.4.1 Définition du problème

Un autre sujet de recherche majeur est le problème de la découverte de règles réellement pertinentes à partir de l'ensemble des règles générées. Ce problème est étroitement lié à la taille de la collection des itemsets fréquents et à la part importante de redondance entre les règles, et par rapport au domaine d'application. Ceci est d'autant plus vrai lorsque les données sont denses ou fortement corrélées [BAG00, BMUT97].

On a vu qu'il était possible de générer un ensemble restreint de règles d'association.

La diminution du volume de règles représente un premier pas pour faciliter l'analyse des résultats. De plus, de nombreuses mesures objectives ont été proposées dans la littérature. Ces mesures permettent d'évaluer certains critères statistiques des règles d'association, permettant ainsi une analyse plus fine de chaque règle, mais, pour autant, elles ne sont pas toujours suffisantes pour sélectionner les règles réellement intéressantes du point de vue de l'utilisateur.

Pour cela, il nous apparaît nécessaire d'introduire des critères subjectifs dans le processus de découverte. Ils peuvent être utilisés dès la phase d'extraction afin de réduire l'espace de recherche. Une autre approche, qui n'est pas indissociable de la première, consiste à effectuer des post-traitements sur les résultats et de sélectionner des règles plus pertinentes vis-à-vis des critères spécifiés par l'utilisateur. On s'intéressera aussi aux mesures d'intérêt dites *subjectives*. De telles mesures peuvent, par exemple, être définies pour prendre en compte une hiérarchie de concepts établie sur le domaine – ou un modèle de connaissance plus élaboré – dans le but de filtrer les règles évidentes, déjà connues, ou encore celles qui ne sont pas utiles pour l'utilisateur.

2.4.2 Post-traitement des règles extraites

Filtrage syntaxique, approches basées sur les patrons

L'approche consistant à utiliser des « patrons » de règles (ou *templates* en anglais) a été formalisée par [KMR+94] dans le cadre du filtrage des règles d'association. Cette approche est utile pour filtrer un ensemble de règles qui ne correspondent pas aux critères définis par un expert. Les « patrons » servent à spécifier ces critères. Un patron de règle est défini par l'expression :

$$A_1, \ldots, A_k \rightarrow A_{k+1}$$

où chaque A est soit un nom d'attribut, un nom de classe, ou une expression C^+ ou C^*, C étant le nom d'une classe. Ici C^+ et C^* correspondent respectivement à « une ou plus »et « zéro ou plus » instances de la classe C.

En pratique cette approche s'utilise très souvent lorsque l'expert des données souhaite guider le processus de découverte de règles. Elle est de plus assez naturelle puisqu'elle se rapproche de requêtes que l'on peut effectuer sur une base de données, ou du filtrage par expressions régulières. Ici, l'expert introduit un biais plus ou moins fort dans les découvertes qu'il va pouvoir réaliser à partir de l'ensemble des règles extraites.

Exploitation des taxonomies

Dans certains cas d'application il peut être intéressant de modéliser certains attributs sous la forme d'une hiérarchie. Un exemple simplifié, tiré de notre cas d'application est présenté dans la figure 2.4.

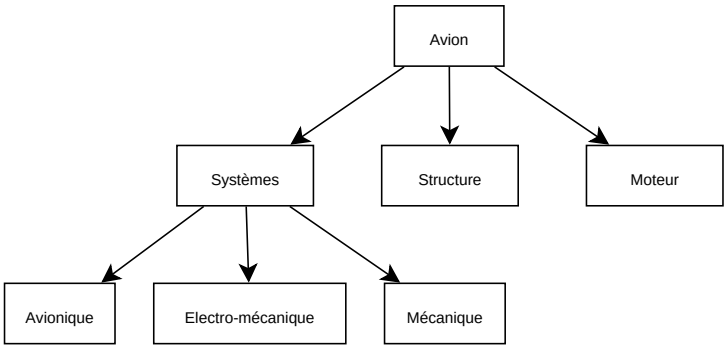

FIG. 2.4 – Exemple simplifié d'une taxonomie présente pour le cas d'application des données IO.

On peut noter qu'au sein d'une hiérarchie les items de niveau inférieur ont des supports inférieurs ou égaux aux niveaux qui leur sont supérieurs. Cet état de fait peut être avantageusement utilisé par les algorithmes d'extraction, mais aussi dans le but d'extraire des règles d'association multi-niveaux. L'intérêt est évident, de cette manière il est possible d'extraire des règles contenant des attributs plus « haut niveau » lorsque les attributs de bas niveau sont peu représentés dans les données. On obtient des règles plus générales mais que la présence d'attributs complémentaires peut rendre intéressantes.

Pour réaliser une extraction multi-niveaux de règles d'association il existe plusieurs types d'approches. La première approche est dite *top-down* avec un seuil de support uniforme pour tous les niveaux. Pour parcourir la hierarchie des items on utilise la propriété de non monotonie de la contrainte de fréquence : si un concept (ou ensemble de concepts) est non fréquent, alors tous ses fils sont non fréquents. Par exemple : si la catégorie *Systèmes* n'est pas fréquente, alors tous ses enfants (i.e. *Avionique*, *Electro-mécanique*, *Mécanique*) ne seront pas fréquents non plus.

Le problème généralement posé par ce type d'approche est que l'on peut rater des règles au niveau inférieur. La littérature est assez abondante sur les optimisations à apporter. On peut par exemple vouloir réaliser une exploration avec un support décroissant, faire l'hypothèse d'indépendance entre les niveaux (du point de vue de la fréquence), utiliser un filtrage sur un item, ou sur k items. Dans tous les cas de figures il faut bien comprendre qu'il y a un surplus de règles générées inutilement, dû à la présence de cette hiérarchie. En effet, certaines règles peuvent être redondantes à cause des relations de « parenté » entre les catégories définies par la taxonomie.

Prenons les deux règles d'association ci-dessous :

$$(1)\text{Systèmes=vrai, Attr2=vrai} \rightarrow \text{Attr3=vrai}[\mathcal{F}_r = 0,06; \text{conf} = 0,7]$$

$$(2)\text{Systèmes=vrai, Attr2=vrai} \rightarrow \text{Electro-mécanique=vrai, Attr3=vrai}[\mathcal{F}_r = 0,02; \text{conf} = 0,72]$$

Dans ce cas, on dit que la règle (1) est un ancêtre de la règle (2). Une règle est redondante si sa valeur de fréquence est très proche du support que l'on pourrait prévoir, en se basant sur la fréquence de la règle ancêtre. Dans notre exemple, si le nœud *Systèmes* a 3 fils, dont *Electro-Mécanique*, alors une fréquence de $0,02$ est « prévisible » (i.e. $0,06/3 = 0,02$) ; ici la règle (2) est jugée redondante par rapport à la (1).

La description d'une taxonomie sur les données est en fait un cas particulier de description des connaissances du domaine d'application. Dans ce cas il s'agit de dépendances « logiques », parfaitement documentées, dont on souhaite tirer parti. Il peut être intéressant, comme nous l'avons formulé pour l'approche par *templates* de réfléchir à une modélisation plus systématique des connaissances. En effet, s'il est utile de pouvoir représenter et exploiter une taxonomie existant sur certains attributs, il est d'autant plus intéressant de modéliser et d'utiliser les relations – quantitatives et qualitatives – sur l'ensemble des différents attributs du domaine.

2.4.3 Extraction sous contraintes

On peut voir APRIORI comme le premier algorithme d'extraction sous contrainte. En effet, il utilise la contrainte de fréquence minimale pour élaguer le treillis des itemsets. On peut facilement généraliser APRIORI afin d'utiliser n'importe quelle contrainte anti-monotone. Ainsi de nombreux algorithmes ont été développés, utilisant des contraintes variées. L'objectif de ces travaux est double : d'une part, optimiser la phase d'extraction de motifs en introduisant des contraintes qui vont réduire l'espace de recherche, et d'autre part permettre à l'utilisateur de spécifier ce qui l'intéresse par le biais de contraintes ou de critères de recherche.

Un exemple d'algorithme utilisant des contraintes autres que la fréquence minimale est fourni par Srikant et al. [SA96], il s'agit de *Direct*. Les auteurs proposent d'extraire les itemsets fréquents qui satisfont une *contrainte syntaxique* donnée par l'utilisateur. Une contrainte syntaxique, est, par définition, une contrainte qui ne dépend pas des données. Soit S un itemset, la contrainte « $A \in S$ est un exemple de contrainte syntaxique. Ces contraintes s'étendent naturellement aux règles d'association. Cela va permettre à l'utilisateur de sélectionner les règles qui sont utiles pour son problème. Il peut, par exemple, préciser qu'il est intéressé uniquement par les règles d'association dont la tête contient un attribut de classe.

Les auteurs de *Direct* introduisent aussi d'autres contraintes dans le cas où il existe une taxonomie sur les items. Rappelons qu'une taxonomie est une relation acyclique qui sert à classifier des items en plusieurs catégories, (un exemple classique est celui de la grande surface qui classe ses produits en différentes catégories, puis sous-catégories, etc.).

L'algorithme *CAP* [NLHP98] propose une approche différente pour pousser des contraintes anti-monotones syntaxiques, ainsi que certaines contraintes utilisant les fonctions d'agrégats (e.g. SOMME(S) < 100).

Ces algorithmes sont intéressants car ils permettent de *pousser* les contraintes, c'est-à-dire de les utiliser pour limiter l'espace de recherche lors de la phase d'extraction des itemsets fréquents.

2.4.4 Conclusion

L'introduction de contraintes spécifiées par l'utilisateur permet de réduire encore un peu plus le nombre de règles à analyser, en se focalisant sur *ce qui intéresse* l'utilisateur, donc en éliminant *les motifs qui ne l'intéressent pas*, ceux qu'il *connaît déjà*.

On distingue alors deux façons de procéder, soit on exprime ces contraintes de manière à ce qu'elles puissent être exploitées au moment de l'extraction, ouvrant ainsi la porte à des bases de données toujours plus complexes ; soit ces contraintes sont utilisées en post-traitement.

Nous pensons que ces approches constituent, en quelque sorte, une façon « déguisée » de décrire et de mettre à profit certaines connaissances du domaine. Il peut être intéressant de voir s'il est possible de généraliser ce type d'approche en introduisant un modèle plus formel des connaissances exprimées par l'expert.

2.5 Comment prendre en compte la connaissance du domaine ?

2.5.1 Définition du problème

On peut effectuer une distinction entre différents types de règles d'association :

1. Une règle d'association peut correspondre à une **connaissance du domaine**, ou une **connaissance attendue**. Ainsi, dans le cas où une base de données enregistre les achats de clients d'une grande surface, on s'attend à voir apparaître un certain nombre de règles décrivant des achats typiques. Comme, par exemple,

la règle suivante :

$$\text{Bière} \rightarrow \text{Chips}$$

La découverte de telles associations, que l'on suppose ici déjà connues, va être nuisible à l'étape d'analyse des résultats car aucune information nouvelle n'est présentée à l'utilisateur.

2. Une règle d'association peut aussi faire référence à des attributs, ou des combinaisons d'attributs, inintéressants. La règle :

$$\text{Pantalon} \rightarrow \text{Chemise}$$

est jugée inutile dans le cas où l'utilisateur ne s'intéresse qu'aux règles contenant au moins un produit « alimentaire ».

3. Les règles peuvent être **redondantes** entre elles.

$$\text{Pantalon, Chausettes} \rightarrow \text{Chemise}$$

$$\text{Pantalon, Chausettes, Caleçon} \rightarrow \text{Chemise}$$

4. Enfin, une règle d'association peut contenir des informations **valides et potentiellement intéressantes**, du point de vue de l'expert, comme la règle Bière → Couches.

Cette première nomenclature des règles d'association met en évidence la nécessité de bien séparer les différents niveaux de traitement qu'il va falloir mettre en place si l'on veut pouvoir converger rapidement vers des règles pertinentes. En effet, si on peut se débarrasser des règles décrivant les connaissances du domaine (ou des connaissances attendues), ainsi que des règles redondantes ou non pertinentes vis-à-vis du contexte, alors il y a de fortes chances pour que les règles restantes soient porteuses d'informations valides et intéressantes.

Nous avons évoqués jusque là différentes techniques qui peuvent se montrer complémentaires pour mener à bien cette tâche. D'une part on sait qu'il est possible de travailler à partir d'un ensemble concis de règles : on élimine ainsi la redondance intrinsèque (règles de la catégorie 3). De plus l'utilisateur a la possibilité de préciser différentes contraintes (filtrage syntaxique, approches basées sur les patrons, exploitation des taxonomies) qui vont lui permettre d'éliminer une part des règles inutiles (catégorie 2). Enfin tout un éventail de mesures objectives va faciliter l'évaluation des règles restantes, à savoir les règles des catégories 1 et 4.

Comme on peut le voir le problème de la redondance au domaine d'application n'a pas encore été résolu.

En appliquant des contraintes spécifiées par l'utilisateur, on fait intervenir un biais dans le processus de découverte de règles pertinentes. Ce biais est essentiel car, sur

des cas d'application complexes les résultats sont trop nombreux et la redondance par rapport au domaine d'application est trop importante. Seulement les méthodes présentées ont toutes pour défaut d'être appliquées de manière *ad hoc*, en utilisant des formalismes divers. On s'intéresse souvent à un aspect précis de ce que recherche l'expert ce qui peut avoir pour effet de trop élaguer l'espace de recherche, ou au contraire de ne pas assez le réduire.

En fait, par le biais de ces contraintes, l'utilisateur du système cherche à modéliser *ce qu'il sait*, de telle sorte que les *connaissances* qu'il exprime ne soient pas présentes dans les règles d'association qu'il va examiner. Cette approche s'inscrit donc dans une démarche d'ingénierie dans le connaissance, dans le sens où l'utilisateur va devoir exprimer ses connaissances afin qu'elles puissent être traduites sous formes de contraintes sur les règles d'associations. Il s'agit en quelque sorte d'aborder le filtrage des règles attendues ou connues de manière plus systématique.

Dans cette section nous allons examiner les principales contributions qui se rapprochent de cette démarche.

Découverte de motifs inattendus

Les auteurs de l'approche *Small is beautiful* [PT00] partent du constat suivant : les techniques de fouille de données traditionnelles n'exploitent pas de manière systématique les connaissances a priori de l'expert. Lorsqu'on dispose –potentiellement– de nombreux experts et analystes qui ont des intuitions sur la problématique étudiée, intuitions basées sur leur expérience, il peut sembler regrettable de ne pas pouvoir exploiter ces ressources.

Ce problème a initié de nombreux travaux de recherche dont le but était la découverte d'un ensemble de motifs qu'on qualifiera d'*inattendus*. La notion d'*inattendu* ou d'*étonnement* ne peut se définir que par rapport à un contexte précis. Dans le cas des travaux de [PT98, PT98, PT06] les auteurs ont trouvé une définition de motifs inattendus par rapport à une croyance exprimée sous forme de règle logique.

Soit une règle A $\rightarrow$ B, cette règle sera jugée « inattendue » par rapport à la croyance $X \rightarrow Y$ si elle respecte les conditions suivantes :

- (a) B ET $Y \models$ FALSE , ce qui signifie que B et Y sont contradictoires logiquement.
- (b) A ET X , cette proposition est vérifiée par un ensemble d'enregistrements suffisants (vis-à-vis d'un critère utilisateur) dans une base de données bd.
- (c) La règle $A, X \rightarrow B$ est vérifiée sur bd.

La clé de l'utilisation de cette définition est l'hypothèse de monotonicité des croyances. En particulier, comme le montrent les auteurs, si la croyance $Y \rightarrow B$ est vérifiée sur un ensemble D, alors la propriété de monotonicité nous dit que cette

croyance sera aussi vérifiée sur tout sur ensemble de D.

En utilisant cette propriété les auteurs mettent au point un algorithme capable d'extraire une collection minimale de règles étonnantes, par rapport à un ensemble de croyances exprimées sous la forme de règles logiques. L'algorithme développé a été validé expérimentalement en comparaison avec l'algorithme APRIORI. Même si la comparaison des performances avec APRIORI est ici anecdotique, puisque comme on l'a dit APRIORI a été conçu pour extraire toutes les règles supérieures à un certain seuil de fréquence, on se rend compte de l'intérêt des approches visant à réduire le nombre de règles extraites. D'une part cela permet de travailler à des seuils de fréquences plus bas, et donc potentiellement plus intéressants aux yeux d'un expert, et d'autre part le nombre de règles à analyser étant plus faible, on tend plus rapidement vers la découverte de règles intéressantes.

On peut cependant formuler un premier problème vis-à-vis de ces travaux. En effet, ici, l'intérêt d'une règle est traité *localement*, ce qui rend impossible la prise en compte de la transitivité. Si l'on prend les règles A $\rightarrow$ B et B $\rightarrow$ C comme représentant les connaissances du domaine, alors la règle A $\rightarrow$ C pourra être jugée intéressante, alors même que les connaissance du domaine nous indiquent, par transitivité, que cette relation est déjà connue.

Un deuxième problème que l'on voit se dessiner pour ce type d'approche est l'étape de définition de l'ensemble des croyances, ou plus généralement des connaissances a priori. Même si l'approche de type « système expert » a été très en vogue a une certaine époque, on sait aujourd'hui qu'il est extrêmement difficile de définir et de gérer une base de connaissance constituée de règles logiques.

Utilisation d'un réseau bayésien pour mesurer le caractère inattendu d'un itemset fréquent

Cette approche a été proposée pour la première fois dans [JS04]. Les auteurs partent du constat suivant : le post-traitement des règles d'association est basé principalement sur l'utilisation de mesures dites *objectives*. Comme précisé précédemment, la majorité de ces mesures évaluent l'intérêt comme une fonction de la divergence entre : la probabilité d'apparition d'une règle calculée sur les données, et sa probabilité d'apparition sous l'hypothèse d'indépendance. Le défaut principal de ce type de mesures est qu'elles ont tendance à faire apparaître des règles déjà connues de l'expert ou évidentes. En effet, les motifs sélectionnés par ces méthodes peuvent être découvert par le biais de méthodes classiques, ou ils peuvent se déduire intuitivement à partir de l'expérience de l'utilisateur.

S. Jaroszewicz et al. pensent que la meilleure façon d'aborder ce problème est de prendre en compte les connaissances du domaine dans le processus de fouille de données. Un motif sera intéressant si il est surprenant pour l'expert, ou « innatendu ».

Le caractère « innatendu » d'un motif (ou en l'occurence d'un itemset fréquent) est déterminé par sa divergence vis-à-vis des connaissances du domaine.

Contrairement à l'approche [PT98] qui consiste à modéliser les connaissances par un ensemble d'implications logiques, puis à utiliser ces connaissances de manière locale, les auteurs de [JS04] proposent de modéliser et d'exploiter la loi jointe de probabilité sur l'ensemble des données. Pour cela ils expérimentent l'utilisation d'un réseau bayésien en tant que modèle des connaissances du domaine.

Avant d'entrer plus en détails dans la description de cette approche qui a, en partie, inspiré nos travaux de thèse, une présentation des Réseaux Bayésiens est nécessaire. Qu'est-ce qu'ils représentent ? Comment sont-ils utilisés ? Comment les modéliser pour qu'ils reflètent avec précision les connaissances de l'expert ?

Ces questions sont essentielles car elles sont au cœur de nos travaux de thèse. Une fois abordées, on reviendra sur l'utilisation des Réseaux Bayésiens, proposée par S. Jaroszewicz et al., dans le cadre de la découverte d'itemsets fréquents inattendus.

2.5.2 Les réseaux bayésiens comme modèle de la connaissance du domaine

Cette section a pour objectif de présenter les *Réseaux Bayésiens* en tant que modèle de connaissance (nous utiliserons l'abréviation RB). Il ne s'agit pas de réaliser, ici, un état de l'art exhaustif sur le domaine, mais bien de mettre en avant les principes, ainsi que les principales contributions qui se rattachent à l'utilisation des Réseaux Bayésiens sur des cas d'applications réels. En particulier on s'intéressera aux limites que présentent leur utilisation lorsque les données envisagées sont complexes (nombreuses variables, elles mêmes décomposées en de multiples valeurs).

Présentation des Réseaux Bayésiens

Un RB est un graphe causal auquel on associe une représentation probabiliste sous-jacente. La circulation de l'information à l'intérieur de ce graphe obéit à des règles très précises, en particulier à la règle de *d-séparation* initialement proposée par J. Pearl en 1988 [Pea88].

Définition 2.20 (d-séparation) *Soit un graphe orienté G composé d'un ensemble de nœuds. Soit X, Y et Z différents nœuds de G. On dira que X et Y sont d-séparés par Z (on notera $\langle X \mid Z \mid Y \rangle$) si pour tous les chemins entre X et Y, l'une au moins des deux conditions suivantes est vérifiée :*

- *Le chemin converge en un nœud W, tel que $W \neq Z$, et W n'est pas une cause directe de Z.*

– *Le chemin passe par Z, et est soi divergent, soit en série au nœud Z.*

Exemple. Les affirmations suivantes illustrent la notion de d-séparation et sont directement déduites de la figure 2.5 :

- $\langle A|B|D\rangle$
  - Le chemin A-B-D est en série en B ($A \rightarrow B \rightarrow D$).
  - Le chemin A-C-D est convergent en C ($A \rightarrow C \leftarrow D$).
- $\langle A|D|E\rangle$
  - Tous les chemins de A à E passent par D.
  - Le chemin A-B-D-E est en série en D ($B \rightarrow D \rightarrow E$).
  - Le chemin A-C-D-E est divergent en D ($C \leftarrow D \rightarrow E$).

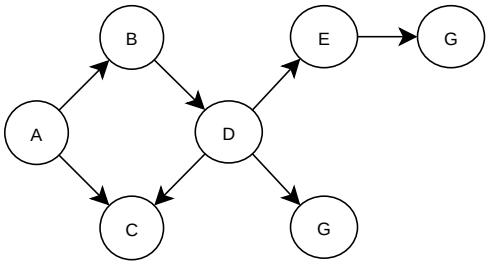

FIG. 2.5 – Exemple de graphe orienté

On notera que la définition de la d-séparation présentée ci-dessus peut être étendue facilement dans le cas où Z est un ensemble de nœuds.

Cette notion, purement graphique, est cependant difficile à appréhender telle quelle. Ainsi, on pourra la formuler de la façon suivante : le fait que X et Y sont d-séparés par Z signifie que Z *bloque* le passage de l'information entre X et Y *dans le cas où* Z est la seule information connue dans le graphe.

A partir de cette propriété, Verma et Pearl ont démontré le deuxième résultat important des RB : « si X et Y sont d-séparés par Z, alors X et Y sont indépendants sachant Z ». Ce résultat est fondamental, il détermine en fait la propriété suivante :

$$\langle X|Z|Y\rangle \rightarrow p(X|Y,Z) = p(X|Z)$$

Ce résultat permet de limiter les calculs de probabilités grâce à des propriétés du graphe. Ainsi, supposons que X et Y soient d-séparés par Z, et que Z soit connun et supposons par ailleurs que l'on vienne de calculer $p(X|Z)$. Si une nouvelle information sur Y est alors connue, le résultat ci-dessus permet de conserver le calcul de $p(X|Z)$ et de le réutiliser comme valeur de $p(X|Z,Y)$. Combiné avec un autre résultat qui établit qu'un nœud est d-séparé du reste du graphe par l'ensemble constitué de ses

parents, de ses enfants, et des autres parents de ses enfants, cette propriété permet de rendre locaux tous les calculs de probabilités dans un graphe causal.

On définit un RB de la façon suivante :

Définition 2.21 (Réseau Bayésien) *$B = (G, \theta)$ est un réseau bayésien si $G = (X, E)$ est un graphe acyclique orienté dont les sommets représentent un ensemble de variables aléatoires $X = \{X_1, \ldots, X_n\}$, et si $\theta_i = [\mathcal{P}(X_i / X_{Pa(X_i)})]$ est la matrice des probabilités conditionnelles du nœud i connaissant l'état de ses parents $Pa(X_i)$ dans G.*

Utilisation et difficultés

A partir des propriétés du RB et de la définition énoncée ci-dessus on peut déduire une propriété importante des RB qui est de définir de manière unique et condensée la distribution de probabilité jointe de H :

$$P_H^{RB} = \prod_{i=1}^{n} P_{A_i|\Pi_{A_i}}$$

Cette propriété va permettre d'effectuer les calculs de probabilités conditionnelles d'événements reliés les uns aux autres par des relations de cause à effet, à l'intérieur du graphe. Cette utilisation du RB s'appelle l'*inférence*.

En termes d'utilisation du modèle, l'avantage essentiel des Réseaux Bayésiens par rapport à d'autres techniques est de permettre une formalisation, assistée par l'expert, des connaissances du domaine, sous forme d'une représentation graphique « lisible ». Néanmoins la construction et la mise au point du réseau bayésien sont considérées comme des problèmes difficiles.

Le problème de l'*apprentissage*, ou *construction* d'un RB doit permettre de répondre à ces deux questions :

- Comment estimer les lois de probabilités conditionnelles ?
- Comment trouver la structure du réseau bayésien ?

Ainsi, on va naturellement diviser le problème de l'apprentissage en deux parties : l'apprentissage des paramètres (avec une structure fixée) et l'apprentissage de la structure.

Inférence dans les Réseaux Bayésiens

Le modèle représenté par le RB n'est pas un modèle statique, fermé. Il est capable d'intégrer de nouvelles informations exogènes. Celles-ci, en modifiant la vraisemblance de certains nœuds, vont modifier les probabilités *a posteriori* de l'ensemble du système. Tout calcul portant sur la distribution de probabilité associée à un RB relève de l'inférence. En fait il ne s'agit pas d'un problème théorique (la distribution de probabilité étant entièrement définie) mais d'un problème de calcul.

Cette opération, l'inférence probabiliste, a été prouvée NP-difficile dans le *cas général* [Coo88]. Pour résoudre ce problème on peut distinguer deux classes d'algorithmes d'inférence : les méthodes exactes et les méthodes approchées.

Dans la catégorie des méthodes exactes on peut faire la distinction entre les méthodes, dites de *propagation des messages* étendues par des algorithmes de coupe (ou de conditionnement) [Pea88] et les méthodes utilisant des groupements de nœuds [LS88], améliorées par la suite par [JJJ96]. Les premières proposent un mécanisme de calcul utilisant la propagation de messages le long des arcs d'un graphe sans cycle, les secondes opèrent d'abord des modifications sur le graphe pour obtenir une structure secondaire d'arbre de jonction dans laquelle chaque nœud représente une clique[5] du réseau bayésien et qui permet d'appliquer un algorithme simplifié de propagation des messages.

Les algorithmes de calcul d'inférence approchée se divisent en deux branches : d'une part les algorithmes qui utilisent des méthodes exactes mais opèrent seulement sur une partie du graphe, d'autre part, les algorithmes qui utilisent des méthodes stochastiques (simulations).

Dans la première catégorie, on retrouvera les contributions de [rul94] qui exploitent le fait que certaines dépendances sont faibles, c'est-à-dire que, qualitativement, il existe un arc entre des nœuds X et Y parce que ces variables ne sont pas exactement indépendantes l'une de l'autre, mais que, quantitativement cette dépendance est insignifiante ; autrement dit X et Y se comportent presque comme si elles étaient indépendantes. L'idée de cet algorithme est donc d'éliminer ce type d'arc afin de simplifier les calculs de propagation des messages, tout en engendrant un taux d'erreur raisonnable.

La deuxième catégorie concerne un ensemble de méthodes qui reposent sur des principes stochastiques. L'idée de départ des méthodes stochastiques est d'utiliser ce que l'on connaît de la loi étudiée pour générer automatiquement des échantillons d'une base de données représentative de cette loi (génération d'exemples). Il suffit ensuite d'utiliser cette base simulée pour calculer les différents estimateurs.

Ainsi, différentes méthodes sont apparues, qui se distinguent par leur façon de me-

[5] On désigne ici par *clique* le graphe induit par un ensemble de sommets deux-à-deux adjacents.

ner les simulations, de générer la base d'exemples en fonction de différentes connaissances de la loi étudiée. On peut citer par exemple les méthodes dites *probabilistic logic sampling*, ou encore les méthodes dites « MCMC » (*Markov Chain Monte Carlo*). Plus précisément, les MCMC sont une famille de méthodes stochastiques comprenant entre autres Metropolis [GRS96] et l'échantillonneur de Gibbs [Nea93].

Pour conclure, on peut dire que l'*inférence* dans les RB est un problème maîtrisé et suivant les cas on pourra faire appel à des méthodes exactes ou approchées, selon que l'on souhaite privilégier la performance (temps de calcul) ou la précision des résultats.

Apprentissage des paramètres, à partir de données complètes

L'estimation de distributions de probabilités (ou les paramètres des lois correspondantes) à partir de données disponibles est un sujet vaste et complexe. On peut conseiller, en particulier, la lecture de [Hec95, Kra98, Jor98].

Dans le cas où toutes les variables sont observées, la méthode la plus simple et la plus utilisée est l'*estimation statistique* qui consiste à estimer la probabilité d'un événement dans la base de données. Cette approche, appelée *maximum de vraisemblance*, nous donne :

$$\hat{p}(X_i = x_k | pa(X_i) = x_j) = \hat{\theta}^{MV}_{i,j,k} = \frac{N_{i,j,k}}{\sum_k N_{i,j,k}} \tag{2.1}$$

où $N_{i,j,k}$ est le nombre d'événements dans la base de données pour lesquels la variable X_i est dans l'état x_k et ses parents dans l'état x_j.

Le lecteur pourra aussi consulter l'approche par *estimation bayésienne* décrite dans [Rob96]. Elle consiste à trouver les paramètres θ les plus probables sachant que les données ont été observées, en utilisant des *a priori* sur les paramètres.

Apprentissage des paramètres, à partir de données incomplètes

Lorsque, ce qui est souvent le cas dans les bases de données réelles, les données sont incomplètes, la problématique d'apprentissage des paramètres est différente. Dans ce cas là, la méthode la plus utilisée repose sur l'algorithme itératif *Expectation-Maximisation* (EM) proposé par A. Dempster et al. en 1977 [DLR77] et appliqué aux Réseaux Bayésiens dans [CDLS03, NH98] ainsi que dans [TMH01].

L'algorithme EM s'applique à la recherche des paramètres du Réseau Bayésien en répétant jusqu'à convergence les deux étapes *Espérance* et *Maximisation*. Il s'utilise aussi bien dans le cadre de l'estimation statistique que pour l'estimation bayésienne.

De plus de nombreux travaux de recherche ont mis en avant différentes heuristiques pour accélérer la convergence de l'algorithme [NH98].

Acquisition des connaissances

Dans de nombreuses applications, réelles, on dispose généralement de très peu de données. Dans ces situations, l'apprentissage des paramètres du RB passe par l'utilisation des connaissances d'experts pour tenter d'estimer les probabilités conditionnelles.

Une première difficulté, souvent appelée *élicitation de probabilités* dans la littérature, et de manière plus générale dans le domaine de l'acquisition de connaissances, consiste à associer une probabilité de réalisation à un « fait » (réalisation d'une variable). Les difficultés liées à cette problématique relèvent généralement de l'ingénierie de la connaissance et de nombreuses méthodes ont été proposées au fil des années. Ci-dessous on détaillera trois types de problèmes spécifiques, ainsi que les solutions que l'on peut trouver dans la littérature :

- le premier problème concerne l'estimation de la probabilité d'un événement par un expert,
- le deuxième concerne l'estimation d'un événement conditionnellement à un grand nombre de variables
- le troisième problème consiste à être capable d'intégrer différentes *sources d'informations multiples*, en tenant en compte de la fiabilité de différents experts et de différentes sources.

De nombreux travaux existent sur l'élicitation de probabilités [Ren01]. La tâche la plus difficile étant de trouver un expert à la fois fiable, disponible, puis de le familiariser à la notion de probabilité. Ensuite il faut tenir compte des biais éventuels, par exemple un expert peut surestimer la probabilité de réussite d'un projet le concernant, etc. La deuxième étape consiste à fournir à l'expert des outils associant des notions qualitatives et quantitatives pour qu'il puisse associer une probabilité aux différents événements. L'outil le plus connu et le plus facile à mettre en place est l'échelle de probabilité, présentée dans la figure 2.6. Cet outil a été introduit par [DG00] et il permet aux experts d'utiliser des informations à la fois textuelles et numériques pour assigner un degré de réalisation à telle ou telle affirmation, puis de comparer les probabilités des événements pour les modifier. [vdGRW+02] présente une étude détaillée des techniques d'élicitation de probabilités.

Le deuxième problème concerne le cas où un expert doit estimer la probabilité conditionnelle $p(Y|X_1, X_2, \ldots, X_n)$. Pour simplifier prenons le cas où toutes les variables (Y et X_i) sont binaires. Dans ce cas l'expert devra estimer 2^n valeurs, ce qui devient rapidement irréaliste dès que n est grand, ce qui est le cas pour de nombreux cas d'applications réels. L'idée est alors de simplifier cette probabilité conditionnelle en posant les hypothèses suivantes :

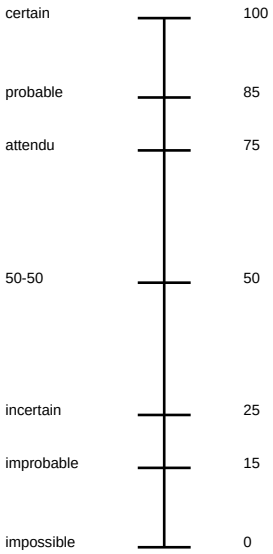

FIG. 2.6 – Échelle de probabilité

– On peut calculer facilement la probabilité suivante : $p_i = p(y|\bar{x}_1, \bar{x}_2, \ldots, x_i, \bar{x}_n)$.
– « X_i cause Y » est indépendant des autres variables X_j.

Le modèle *OU bruité* nous dit alors que :

– Si un des X_i est vrai, alors Y est presque toujours vrai (avec la probabilité p_i),
– Si plusieurs X_i sont vrais, alors la probabilité que Y soit vrai est donnée par :

$$p(y|\mathcal{X}) = 1 - \prod_{i|X_i \in \mathcal{X}_p} (1 - p_i) \tag{2.2}$$

où $\mathcal{X}_p$ est l'ensemble des X_i vrais.

Ce modèle proposé initialement par J. Pearl [Pea88], a été étendu au cas où Y peut être vrai sans qu'une seule des causes soit vraie (*leaky noisy-OR gate*) et aux variables multi-valuées (*generalized noisy-OR gate*).

Cette modélisation simplifiée des probabilités a été utilisée avec succès dans des domaines tels que le diagnostic médical [PPMH94, ODW00] ou le diagnostic de pannes [BRM02].

Enfin, le dernier type de problème lié à l'ingénierie des connaissances, concerne le cas où l'ingénieur doit faire face à des sources d'informations de diverses natures : experts, données collectées selon des moyens variés, etc. La prise en compte de ces différentes sources doit se faire avec précaution pour éviter d'utiliser des données biaisées. Ainsi, [DD00] proposent un critère pour vérifier si les différentes sources d'informations ont été utilisées dans les mêmes conditions. Supposons maintenant que

plusieurs experts proposent une estimation sur les mêmes valeurs. Comment combiner ces différents résultats ? La prise en compte de données incertaines a été abordée sous différents angles : la logique floue [BMM03], ou la théorie des croyances [Sme00]. Les auteurs de [PuG+02] proposent quant à eux une méthode qui permet de combiner l'estimation des probabilités faite par un expert avec celle obtenue à partir des données.

Apprentissage de la structure

Comment trouver la structure de réseau bayésien qui représentera le mieux le problème ? L'apprentissage de la structure à partir des données est aussi considéré comme étant un problème NP-difficile. Une première approche consiste à rechercher les différentes relations causales qui existent entre les variables. Les autres approches essaient de quantifier l'adéquation d'un réseau bayésien au problème à résoudre, c'est à dire en associant un score à chaque structure de RB étudiée. Puis elles recherchent la structure qui donnera le meilleur score dans l'espace de recherche des graphes acycliques orientés.

Une approche exhaustive est impossible en pratique en raison de la taille de l'espace de recherche, en effet le nombre de structures possibles à partir de n nœuds est super-exponentiel. Par exemple pour 10 nœuds le nombre de structures possibles est de $4,2 \times 10^{18}$.

Pour résoudre ce problème, un certain nombre d'heuristiques ont été proposées. Elles ont pour but de restreindre cet espace à l'espace des arbres (MWST), d'ordonner les nœuds pour limiter la recherche des parents possibles pour chaque variable (K2), ou d'effectuer une recherche gloutonne (GS).

En partant du principe que plusieurs structures encodent la même loi de probabilité (équivalence de Markov) et possèdent le même score, d'autres méthodes proposent de parcourir l'espace des équivalents de Markov (espace toujours super-exponentiel mais possédant de meilleures propriétés).

Enfin, il existe aussi des méthodes permettant d'incorporer des connaissances a priori sur le problème résoudre en détaillant le plus possible les contraintes que l'on souhaite formuler sur l'espace de recherche.

La littérature concernant cette problématique est trop abondante pour figurer ici, ainsi, pour une lecture plus détaillée de ces différentes approches on pourra se reporter à [NWL+04], ainsi qu'à [FL04] pour une comparaison synthétiques des algorithmes d'apprentissage de la structure.

Il peut par contre être important de retenir que les techniques d'apprentissage automatique bien qu'intéressantes ne peuvent pas être utilisées telles quelles, les résultats obtenus nécessitent impérativement d'être validés par un expert. En effet, on se rend

compte que, même sur des cas d'applications relativement simples, chaque méthode obtient des résultats de structure légèrement différents. Malheureusement ce qui, en termes de *distance d'édition* ne représente qu'une différence singulière (présence ou non d'un arc, inversion du sens d'un arc) peut avoir, pour l'expert du domaine, un sens bien plus critique. De plus, la structure qui satisfait un critère de score (vraisemblance de la structure par rapport aux données d'apprentissage) ne représente pas forcément dans la réalité les connaissances que souhaite modéliser l'expert.

Incorporation des connaissances pour l'apprentissage de la structure

Dans la plupart des cas d'applications, les connaissances de l'expert sur la structure que peut avoir le RB ne sont que partielles. [CGK+02, NWL+04] ont fait une liste de ces connaissances a priori :

1. Déclaration d'un nœud racine (sans parents).
2. Déclaration d'un nœud feuille (sans enfants).
3. Existence (ou absence) d'un arc entre deux nœuds précis.
4. Indépendance de deux nœud conditionnellement à certains autres.
5. Déclaration d'un ordre (partiel ou complet) sur les variables.
6. Déclaration d'un nœud « cible » (pour les applications de type « classification »).
7. Existence d'une variable latente entre deux nœuds.

Quel que soit le type de connaissances apportées par l'expert, il est souvent nécessaire d'utiliser des données pour initier la structure du RB. Les a priori de (1) à (5) peuvent être facilement intégrés aux algorithmes d'apprentissage de structure basés sur l'optimisation d'un score. Les points (6) et (7) font l'objet d'une étude plus spécifique présentée dans [NWL+04] (section 6.2.4).

2.6 Exploitation des Réseaux Bayésiens pour mesurer l'intérêt d'ensembles d'attributs fréquents

Nous avons vu comment construire et exploiter le RB en tant que modèle des connaissances du domaine, revenons maintenant aux travaux présentés dans [JS05] et [JS04].

L'approche envisagée par les auteurs repose sur l'estimation de la fréquence des itemsets à partir du RB et la comparaison de cette estimation avec la fréquence constatée sur le jeu de données. Les itemsets dont la fréquence estimée diverge fortement de la fréquence constatée sont considérés comme intéressants.

Soit bd une base de données binaire de schéma $< T_{id}, \mathcal{I} >$. Les valeurs possibles pour les attributs sont désignés par des lettres minuscules correspondants aux attributs. Notons P_I la distribution de probabilité jointe de l'ensemble d'attributs I. De même notons $P_{I|J}$ la distribution de probabilité de I conditionnée par J. La notation $P_I(i)$ désigne la probabilité pour que $I = i$. La distribution de probabilité estimée à partir des données sera notée en ajoutant le symbole du chapeau, par exemple $\hat{P}_I$. Ainsi la fréquence d'un itemset (I, i) s'écrit :

$$\mathcal{F}(I, i) = \hat{P}_I(i)$$

.

Prenons maintenant un réseau bayésien RB sur un ensemble d'attributs H et un graphe G, on rappelle que RB définit de manière unique la distribution de probabilité jointe de H :

$$P_H^{BN} = \Pi_{i=1}^{n} P_{A_i|\text{par}_I}$$

Avec par_I l'ensemble des attributs parents directs de A_i dans G.

Afin de calculer la mesure d'intérêt d'un ensemble d'attribut, on défini la fréquence d'un itemset (I, i) calculé par rapport au RB, de la manière suivante :

$$\mathcal{F}_{RB}(I, i) = P_I^{RB}(i)$$

L'intérêt d'un itemset (I, i) par rapport à RB, est alors donné par la différence en valeur absolue entre la fréquence de l'itemset constatée sur les données et celle estimée à partir de RB :

$$\text{Int}_{RB}(I, i) = |\mathcal{F}(I, i) - \mathcal{F}_{RB}(I, i)|$$

Les auteurs pensent que les motifs qui ne suggèrent pas une « direction »d'influence sont les plus appropriés dans un contexte de fouille. Ainsi ils s'intéressent uniquement à l'intérêt des itemsets fréquents, voire d'ensemble d'attributs, puisque l'encodage du RB dépend d'attributs et non d'itemsets. De ce fait, ils définissent l'intérêt d'un ensemble d'attributs I de la façon suivante :

$$\text{Int}(I) = \max_{ß \in \text{Dom}(I)} \text{Int}(I, i)$$

Cette mesure d'intérêt est ensuite utilisée pour filtrer et trier les ensembles d'attributs fréquents, afin de faciliter la lecture des résultats : un ensemble d'attribut I est jugé ϵ-intéressant si sa valeur d'intérêt par rapport au RB est supérieure au seuil ϵ. Cependant les auteurs ont constaté que le nombre de motifs ayant un intérêt élevé était trop important. Ainsi, ils ont envisagé deux contraintes afin d'élaguer plus finement l'espace des attributs ϵ-intéressants.

La première contrainte est une contrainte hiérarchique. Elle nous dit qu'un ensemble d'attributs est hiérarchiquement ϵ-intéressant si aucun de ses sous-ensembles n'est ϵ-intéressant.

La deuxième contrainte tire quant à elle partie de la topologie du graphe associé au RB. Un ensemble d'attributs I sera topologiquement ϵ-intéressant si I est ϵ-intéressant et s'il n'existe pas d'ensemble d'attributs J tels que :

- $J \subseteq \text{anc}(I) \cup I$, et,
- $I \nsubseteq J$, et,
- J est ϵ-intéressant.

$\text{Anc}(I)$ est l'ensemble des attributs ancêtres de I dans le graphe G. Cette contrainte permet donc de limiter le fait que la topologie du graphe entraîne une cascade d'attributs intéressants à partir d'un seul attribut intéressant.

Ces contraintes sont alors appliquées par un algorithme de type APRIORI et illustrées sur un jeu de données exemple tiré du répertoire *UCI Machine Learning* [AA07]. Les résultats présentés dans [JS04] sont prometteurs, mais restent cependant peu détaillés. Un premier RB est modélisé par une personne non-experte. L'algorithme calcule ensuite les ensembles d'attributs ϵ-intéressants. Ces découvertes sont alors utilisées pour apporter des modifications manuelles au RB (structure et/ou paramètres). Une modification est validée si le score du RB modifié est supérieur au score du RB précédent. Mais ce score est purement objectif et ne fait que mesurer la probabilité attendue d'avoir les données à partir de la structure : il ne donne aucune indication sur une meilleure adéquation du réseau par rapport aux connaissances expertes du domaine.

En l'absence de cas d'application concret les auteurs semblent privilégier l'amélioration des temps de calculs de leur approche [JS05].

2.6.1 Conclusion

On a vu qu'il existait un ensemble d'approches visant à éliminer les règles inintéressantes du point de vu de l'expert. En effet, une grande partie des règles d'association générées contiennent des informations déjà connues, prévisibles, inintéressantes, ou redondantes. Ces techniques ont été mises en place pour permettre à l'utilisateur de formuler explicitement ce qu'il cherche à découvrir, ou à ne pas découvrir.

Cependant, l'utilisation de chacune de ces méthodes se fait de manière *ad hoc*, ce qui rend d'autant plus difficile leur réutilisation sur de nouveaux jeux de données. De plus, on crée un biais relativement important quant à l'espace qui va être élagué, ce qui réduit l'intérêt d'une approche fouille de données non supervisée. Enfin, on se rend compte qu'il manque une réflexion globale sur la modélisation et la prise en compte des comptes des connaissances du domaine pour faciliter cette étape de découverte.

Les approches présentées [PT98, JS04] sont un premier pas vers une utilisation plus systématique des connaissances de l'expert, dans le but de faciliter la découverte de règles inattendues (et donc potentiellement intéressantes).

Il apparaît important de réfléchir à un processus d'extraction qui prendrait en compte les connaissances du domaine et permettrait de visualiser les règles d'association qui apparaissent inattendues, vis à vis de ce qui a été modélisé. Cette approche doit pouvoir englober aussi bien le traitement des taxonomies (ou des implications logiques) au sein des données, que des connaissances plus fines de l'expert.

2.7 Discussion sur l'état de l'art

Cet état de l'art fait apparaître une évolution des approches pour la découverte de règles d'association. Nous avons commencé par étudier les algorithmes d'extraction d'itemsets fréquents. Les propositions actuelles sont performantes même dans les cas où les données sont fortement corrélées et font intervenir de nombreux attributs. Même si cet axe de recherche est toujours actif, les propositions actuelles s'attachent plus à l'optimisation de leurs algorithmes qu'à une remise en cause des principes utilisés pour l'extraction.

Vient ensuite le problème de la génération des règles d'association et leur analyse par un expert. Nous avons vu qu'il était possible de générer des ensembles non redondants de règles [Pas00], mais aussi qu'on disposait d'un ensemble de mesures objectives [HH99] qui facilitent l'étude des règles. En pratique cependant ces mesures ne sont pas toujours suffisantes pour garantir la découverte de règles réellement intéressantes pour l'expert du domaine.

D'autres approches ont alors montré qu'il était possible d'introduire une part de subjectivité plus importante, notamment par la définition et l'exploitation de taxonomies du domaine [SA96, HMWG98, SCH05], ou encore par le biais de contraintes syntaxiques. Ces approches constituent un premier pas vers l'utilisation de connaissances, en intégrant le jugement de l'expert au processus d'extraction de règles. Cependant, nous sommes convaincus que cette manière de procéder connaît rapidement des limites. Ainsi la redondance par rapport au domaine d'application n'est pas éliminée de manière systématique, il n'y a pas toujours de réelle amélioration apportée à la phase d'analyse des règles.

Concernant l'élimination de la redondance vis-à-vis des connaissances du domaine, nous avons décrit plusieurs approches qui ont en commun une étape de modélisation, puis d'exploitation d'un modèle de la connaissance du domaine. Notamment nous avons vu l'utilisation de réseaux bayésiens pour mesurer l'intérêt d'itemsets fréquents. Cette dernière proposition ouvre des perspectives intéressantes sur la collaboration entre réseau bayésien et règles d'association, perspectives que nous avons décidé d'ap-

profondir dans ces travaux de thèse.

À notre connaissance, il n'y a pas encore de réflexion sur la complémentarité de ces différentes approches. Il s'agit là d'un réel manque pour la littérature sur le domaine, qui donne le sentiment que les solutions proposées sont *isolées* et s'occupent d'une catégorie de problèmes bien spécifiques. Nos travaux se sont donc attachés à décrire la complémentarité de différentes techniques et outils qu'il est nécessaire de mettre en œuvre lorsque l'on s'intéresse à la découverte de règles d'association réellement intéressantes, sur des domaines et des données relativement complexes.

Chapitre 3

Le travail de recherche

Dans nos travaux de thèse nous avons envisagé la mise en place d'un processus de découverte de connaissances, à partir de l'extraction et de l'étude de règles d'association aux propriétés particulières. Les difficultés fréquemment rencontrées lors de la phase d'analyse nous ont amené à réfléchir aux méthodes et techniques nécessaires au déroulement optimal de cette étape, cruciale au sein du processus de découverte de connaissances. Nous nous sommes notamment intéressés à la définition, l'exploitation et l'évolution d'un réseau bayésien comme modèle des principales dépendances du domaine.

3.1 Positionnement par rapport à l'état de l'art, contributions envisagées

Concernant la découverte de règles d'association pertinentes, différentes approches de la littérature ont été détaillées au Chapitre 2. Les problèmes liés aux performances des extracteurs, ainsi qu'à la compacité des représentations obtenues (i.e., et donc indirectement à la redondance des règles présentées) ont été résolus. Parmi les problématiques de recherches demeurant « ouvertes » nous avons choisi d'aborder les axes suivants :

1. La génération d'un ensemble concis de règles à partir des itemsets δ-libres fréquents et de la δ-fermeture.
2. L'exploitation d'un réseau bayésien pour faciliter l'analyse et la découverte de règles d'association pertinentes.
3. La prise en compte des problématiques issues de l'ingénierie de la connaissance au sein d'un processus de découverte de connaissances : définition du modèle initial, annotation des règles d'association, évolution du modèle des dépendances

du domaine.

Les paragraphes qui suivent détaillent notre positionnement par rapport à l'état de l'art. Puis, dans la suite de ce chapitre, nous développons notre travail de recherche pour finir par la validation expérimentale de nos contributions sur le domaine *Visit Asia*.

Génération d'un ensemble concis de règles d'association à partir des δ-libres et de la δ-fermeture

Cette première contribution est inspirée des travaux de recherche de [Pas00, CB02, BBR03]. Ils ont étudié les propriétés des représentations condensées utilisant les itemsets (δ-)libres fréquents et l'opérateur de fermeture. On a aussi vu qu'il était possible de générer un ensemble concis de règles d'association à partir d'une représentation utilisant les libres. Un autre axe de recherche considérait quant à lui l'utilisation de règles dites δ-fortes (corps minimal, confiance contrôlée) dans le cadre de la classification.

Dans nos travaux nous avons envisagé une généralisation de ces différentes approches. En effet, nous utilisons un algorithme [BBR00] capable d'extraire une représentation condensée utilisant les itemsets δ-libres fréquents et la δ-fermeture. Si le comportement est connu lorsque δ est égal à zéro, il est intéressant d'étudier les propriétés de ces itemsets et des règles générées lorsque δ est strictement supérieur à zéro.

Découverte de règles d'association pertinentes par l'exploitation d'un réseau bayésien

Ce deuxième axe de contribution vient du manque actuel – constaté par notre étude sur l'état de l'art – de techniques génériques visant à faciliter la phase d'analyse des règles d'associations. En fait, parmi les solutions actuellement proposées, beaucoup sont trop spécifiques (filtrage à base de patrons, exploitation de la taxonomie, ...) et ne remontent pas à la source du problème qui pose la question suivante : comment utiliser de manière judicieuse les connaissance du domaine pour la phase d'analyse des nombreuses règles extraites ?

La proposition que nous avons faite dans [FDMB06a] consiste à intégrer la connaissance sur les principales dépendances du domaine, au calcul de l'intérêt des règles d'association. L'expert modélise les connaissances qui vont lui servir à éliminer les motifs connus. Il utilise pour cela le formalisme des Réseaux Bayésiens. Les dépendances modélisées permettent de filtrer les motifs témoins dans les données de ces dépendances et facilite ainsi la découverte de règles plus pertinentes.

Une approche similaire [JS04, JS05] a inspiré nos travaux de recherche. Les auteurs

ont montré que les Réseaux Bayésiens étaient un bon support pour modéliser et exploiter les connaissances du domaine, dans le but de favoriser la découverte de motifs divergents par rapport à ce modèle.

Nos travaux se différencient de ceux de S. Jaroszewicz et al. sur plusieurs points :

- L'étude des mesures d'intérêt sur les règles d'association et en particulier sur les règles générées à partir d'une représentation condensée utilisant les δ-libres. Les articles présentés jusqu'à présent, ont choisi de se limiter à la découverte d'ensembles d'attributs intéressants.
- L'exploitation de la structure du RB (i.e. les (in)dépendances graphiques entre les variables) pour la décomposition d'une règle d'association en sous-parties – que nous appelerons motifs– qui reflètent respectivement ce qui est modélisé par le RB et ce qui ne l'est pas. Ce point n'a pas été abordé par les auteurs des propositions sur l'utilisation des RB pour la fouille de règles d'association, mais nous pensons que l'utilisation *explicite* de la propriété de d-séparation peut permettre une décomposition plus fine de l'information portée par les règles.
- La définition et l'évolution du RB au cours du processus de découverte de connaissances. Sur ce point aussi les articles actuels sont relativement évasifs : il a été montré qu'il était possible, par éditions successives du RB, de converger vers un maximum local de l'indice de performance du RB (e.g., rapport à la distribution de probabilité calculée sur les données et celle induite par la configuration du réseau). Cependant, le critère de convergence utilisé est purement objectif et les modifications du réseau ne reflètent pas forcément l'évolution des connaissances du domaine. D'autre part, le processus de mise à jour du RB n'est ni encadré, ni facilité ; ce qui implique de nombreux essais avant de converger vers une solution locale.
- Une réflexion générale sur le processus de découverte de règles pertinentes, le rôle de l'expert, la modélisation du RB initial et le suivi de son évolution, ainsi que sur les outils à implémenter pour encadrer ce processus et faciliter l'analyse des règles. Cet point de vue n'a pas non plus été abordé par les auteurs de ces travaux.
- Enfin, une expérimentation et une validation de nos contributions sur des données réelles. Les travaux de Jaroszewicz et al. étant pour l'instant uniquement présentés sur des données simulées.

Définition d'un ensemble d'outils et de méthodes pour accompagner le processus de découverte de connaissances

La principale originalité de nos travaux, par rapport aux pratiques actuelles en ECD – et plus particulièrement dans le domaine de la découverte de règles d'association pertinentes – est que nous avons inscrit nos différentes contributions dans une approche d'ingénierie de la connaissance.

Ainsi, plutôt que d'imaginer un réseau bayésien fixé a priori, nous avons envisagé l'étude de la construction itérative de ce RB. L'idée que nous avons développée est la suivante : une amélioration du modèle à l'itération i du processus apportera une aide à l'expert dans la phase d'analyse des règles, à l'itération $i+1$. Les informations présentées par les règles sont alors un peu plus pertinentes, c'est-à-dire, plus en phase avec les intentions de l'expert et plus surprenantes par rapport aux connaissances a priori. Une annotation structurée de ces informations permet de capturer de nouvelles dépendances et ainsi de continuer à améliorer le modèle avant de passer à une nouvelle itération de notre processus. A l'étape $i+2$, la pertinence des règles présentées est assurée par les modifications apportées au modèle, et l'utilisation des annotations précédemment collectées. Ainsi, notre processus permet les améliorations successives d'un modèle des dépendances du domaine. Il s'agit en quelque sorte de l'application d'un cercle vertueux pour la production de règles toujours plus pertinentes, grâce à un modèle capturant de mieux en mieux les dépendances du domaine et une collection d'annotations toujours plus riches.

3.2 Proposition de processus de découverte de connaissances : l'approche KARD

Notre première contribution de recherche est partie du constat suivant : les approches actuelles pour la découverte de connaissances n'exploitent pas, ou peu, les connaissances existantes sur le domaine d'application, quelles soient présentes de manière implicite ou non.

Les difficultés liées à l'utilisation de ces modèles s'articulent principalement autour de trois axes :

- La définition du modèle en lui même est généralement considérée comme une entreprise lourde, nécessitant des moyens importants. Il pourra donc être intéressant de trouver un compromis entre la précision du modèle et les ressources déployées pour sa création.
- La définition de mesures et d'outils permettant l'exploitation de ce modèle au sein d'un processus de découverte ; dans le but de sélectionner des règles pertinentes, c'est-à-dire des règles qui présentent une divergence par rapport au modèle.
- Et enfin, la prise en compte des problématiques d'évolution, et de la maintenance du modèle au cours du temps.

Notre contribution s'articule autour de ces trois points. L'approche proposée s'intitule KARD pour *Knowledge-driven Association Rules Discovery*, elle place l'expert et les connaissances du domaine au cœur de la problématique de découverte de règles. L'idée est de réunir deux domaines que nous pensons complémentaires : celui de l'ingénierie de la connaissance et celui de la fouille de données. Les sections suivantes

décrivent les principales étapes de notre approche, avant de détailler le processus général.

3.2.1 Présentation de notre approche

D'un point purement applicatif le processus de *découverte de connaissances* que nous proposons peut être représenté par la figure 3.1. Cette figure permet de visualiser les *entrées* (partie gauche du schéma), les *contrôles* (partie inférieure ou supérieure) et les *sorties* (partie droite) d'une activité donnée (boîte centrale). En l'occurrence nous avons choisi d'orienter notre étude selon le contexte suivante : à partir d'un ensemble de données, définir un processus de découverte qui aboutit à la construction d'un modèle de connaissance (le réseau bayésien) et à la découverte d'une collection $\mathcal{L}$ de règles d'association pertinentes aux yeux de l'expert. Ce processus est contrôlé par une problématique spécifique, les connaissances ainsi que l'expertise disponible sur le domaine d'application.

Le réseau bayésien créé au cours du processus n'est pas l'objectif principal de notre processus. Cependant il peut être considéré comme un effet de bord intéressant : d'une part il modélise un ensemble de dépendances du domaine d'application et d'autre part il pourra être réutilisé sur des problématiques similaires lorsque, par exemple, les données du système évoluent, ou lorsque la problématique est modifiée.

L'évolution de la *connaissance* du domaine (c'est-à-dire l'état des connaissances avant et après le déroulement du processus de découverte de connaissances), va déterminer l'efficacité de notre approche. Comme il nous est impossible d'établir une mesure précise des connaissances du domaine à un instant t, on se fiera plutôt à l'impact *potentiel* que peuvent avoir les motifs découverts, d'un point de vue opérationnel. La question qu'il faut se poser est donc la suivante : les motifs découverts sont-ils d'une quelconque utilité ? Ainsi, si l'expert reconnaît qu'une règle ou qu'un ensemble de règles lui est bénéfique dans son activité (que ce soit en relation directe avec la problématique initiale ou non), alors on pourra dire que la connaissance du domaine est *augmentée*.

On remarque la présence d'une boucle sur une des entrées de notre processus. Ce point sera évidemment explicité dans ce chapitre. Sur le principe il faut juste comprendre que notre approche est itérative : on démarre le processus de découverte de connaissances à partir d'un modèle initial (le réseau bayésien RB_0) ; puis à l'issue de chaque itération on obtient un nouveau modèle (RB_i) qui sera réutilisé à la place du modèle RB_$(i-1)$. Le but de nos travaux est de montrer que cette boucle permet à la fois de converger vers l'élaboration d'un modèle *augmenté* des connaissances sur les dépendances du domaine, mais aussi vers la découverte de règles d'association de plus en plus pertinentes aux yeux de l'expert.

On notera aussi qu'on suppose la base de données d'entrée comme étant préalable-

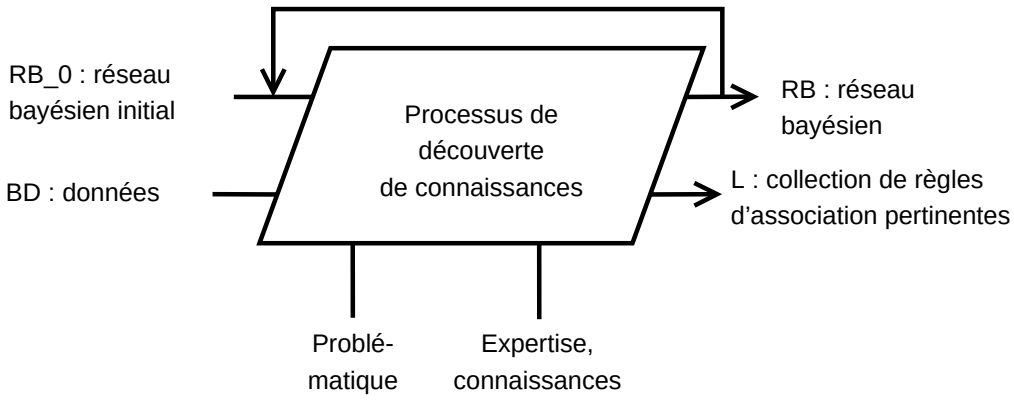

FIG. 3.1 – Activité « Processus de découverte de connaissances ».

ment consolidée et préparée pour les algorithmes d'extraction de règles d'association. Ces différents pré-traitements ne constituant pas l'objet de nos travaux ils ne seront pas détaillés ici.

Cette activité se décompose en plusieurs sous-activités élémentaires que nous détaillons dans les paragraphes suivants ; la première d'entre elles consiste en l'initialisation de notre modèle RB_0. Enfin le cycle complet regroupant toutes les activités sera présenté à la fin de cette section.

Une modélisation des dépendances du domaine

L'approche que nous proposons a pour objectif la découverte de règles qui se révèlent intéressantes *au vu des connaissances du domaine*. L'étude sur l'état de l'art a montré qu'une partie des contributions actuelles visait à introduire la subjectivité nécessaire à la découverte de ces règles, par le biais de contraintes spécifiées par l'utilisateur. D'autres approches ont présenté une vision plus *systématique* de l'exploitation de ces connaissances, en introduisant une modélisation préalable des connaissances.

Les problématiques de *modélisation de la connaissance* présentent évidemment de nombreuses difficultés. Il faut être en mesure de trouver un consensus entre les différents experts du domaine et mobiliser des équipes pour la formalisation et la construction du modèle. Il s'agit d'une activité qui demande généralement un investissement important en termes de temps et d'énergie dépensée. De plus, les bénéfices réels que l'on pourra retirer de l'utilisation du modèle s'avèrent difficile à évaluer a priori, ce qui rend le choix de s'investir dans une telle entreprise à la fois *stratégique* et *risqué*.

Pour ces raisons on préférera utiliser le terme de *dépendances du domaine* plutôt que de parler directement de *modèle des connaissances*, ou encore d'*ontologie du*

domaine. Le réseau bayésien n'est pas pour autant considéré comme une « sous » approche pour la modélisation des connaissances – sa définition requiert du temps, de l'expertise et elle doit faire face à de nombreux problèmes d'ingénierie de la connaissance. Dans notre cas il s'agit plus d'étudier une catégorie de modèle que nous pensons comme étant la plus adaptée pour l'analyse de règles d'association, que d'exploiter le modèle de connaissances *le plus complet* possible sur le domaine.

La figure 3.2 montre les entrées/sorties ainsi que les contrôles liés à l'activité qui consiste à modéliser un premier ensemble des dépendances du domaine.

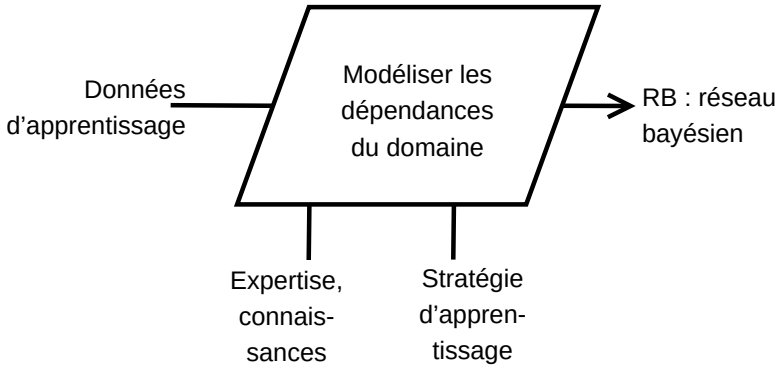

FIG. 3.2 – Activité « Modéliser les dépendances du domaine ».

Un point intéressant à noter ici, est la notion de *stratégie d'apprentissage* envisagée. Ainsi, si nous avons choisi d'intégrer des connaissances du domaine au processus de découverte, notre approche se distingue de la littérature par le fait que la modélisation des connaissances n'est pas figée. Le modèle initial évolue au cours du processus, en bénéficiant des interactions de l'expert sur les motifs extraits. De ce fait, l'étape de définition d'un réseau bayésien initial ne demande pas un investissement initial *démesuré* pour l'expert : en pratique, il va lui être demandé de décrire les *principales* dépendances entre les variables qui définissent le domaine, de manière qualitative et quantitative. Il s'agit là d'une particularité essentielle de notre approche : l'objectif étant de réduire au minimum le coût de construction initial, puis de répartir l'évolution du modèle sur les différentes itérations de notre processus de fouille. Ce faisant, on introduit cependant une autre problématique à notre contexte, celle de la mise à jour du réseau bayésien.

Extraction de motifs locaux

Les règles d'association sont au centre des interactions de notre système. Elles sont extraites à partir d'une base de données bd de schéma $\langle T_{id}, Items \rangle$, en spécifiant un seuil de fréquence minimal $minfreq$, un paramètre δ qui limite la confiance des règles générées au seuil $minconf$, et, éventuellement, un ensemble $\mathcal{C}_S$ de contraintes

syntaxiques. Le résultat de la fonction d'extraction que nous appellerons

$$\phi(\mathrm{bd}, minfreq, \delta, \mathcal{C}_S),$$

et du calcul de n mesures objectives, est la collection

$$\mathcal{L}\langle \mathcal{R}, \mathcal{I}_{\mathrm{bd}}, I_{\mathrm{RB}}, \mathcal{D}\text{-sep}\rangle.$$

$\mathcal{L}.\mathcal{R}$ est l'ensemble des règles d'association qui satisfont les contraintes de fréquence minimale, de confiance minimale, et de syntaxe. De plus, pour toute règle d'association $R_k \in \mathcal{L}.\mathcal{R}$, il n'existe pas de règle d'association $R' \in \mathcal{L}.\mathcal{R}$ telle que R' soit plus générale que R_k.

$\mathcal{L}.\mathcal{I}_{\mathrm{bd}}$ est l'ensemble des mesures associées aux règles d'association de $\mathcal{L}.\mathcal{R}$. À chaque règle R_k on associe n mesures objectives calculées sur bd. On accède à la $i^{\text{ème}}$ mesure d'intérêt de la $k^{\text{ème}}$ règle de la façon suivante : $\mathcal{I}^i_{\mathrm{bd}}(k)$.

Cette activité est représentée par la figure 3.3.

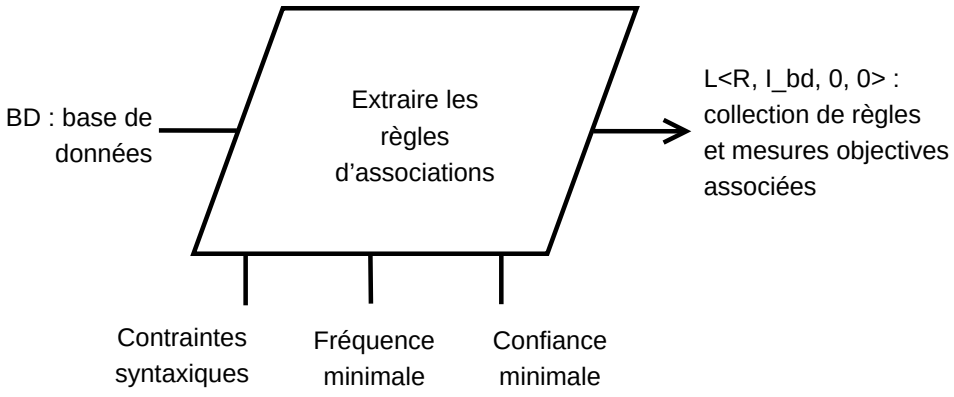

FIG. 3.3 – Activité « Extraire les règles d'associations ».

L'extraction se déroule en deux étapes. La première consiste à extraire une représentation condensée des itemsets fréquents. L'outil que nous utilisons pour cette tâche est AC-like, développé par J. Besson, il s'agit d'une implémentation de l'algorithme MIN-EX. Cet algorithme calcule une représentation utilisant les itemsets δ-libres et leur fermeture.

La deuxième étape repose sur une de nos contributions, à savoir la génération de $\mathcal{L}.\mathcal{R}$, ensemble non-redondant de règles d'association, à partir de ce type de l'ensemble des δ-libres fréquents et de leur fermeture.

Nous calculons *a posteriori* les contraintes syntaxiques, ainsi qu'un ensemble de mesures d'intérêts objectives qui serviront lors de l'étape d'analyse des résultats. Nous avons évoqué dans l'état des l'art des algorithmes capables de *pousser* les contraintes

syntaxiques, c'est-à-dire de n'extraire que les règles qui se conforment aux contraintes. Pour notre application cette optimisation ne s'est pas révélée cruciale.

Exploitation des dépendances du domaine

On cherche à sélectionner des règles pertinentes pour l'expert du domaine. L'« information » portée par les règles d'association extraites est donc « comparée » à celle que l'on peut inférer du réseau bayésien RB_i défini pour l'itération en cours. Pour cela on définit la fonction suivante :

$$\rho(\mathcal{L}\langle\mathcal{R}, \mathcal{I}_{\text{bd}}, \emptyset, \emptyset\rangle, \text{RB}_i, \epsilon) \rightarrow \mathcal{L}'\langle\mathcal{R}', \mathcal{I}'_{\text{bd}}, \mathcal{I}_{\text{rb}_i}, \mathcal{D}\text{-sep}\rangle$$

où ϵ est le seuil d'intérêt subjectif par rapport au réseau bayésien RB_i.

Elle retourne une nouvelle collection $\mathcal{L}'$ telle que :
- $\mathcal{R}' \subseteq \mathcal{R}$, est l'ensemble de règles d'association qui satisfont à la fois les critères objectifs et le critère subjectif ϵ.
- $\mathcal{I}'_{\text{bd}} \subseteq \mathcal{I}_{\text{bd}}$, est l'ensemble des mesures objectives associées à $\mathcal{R}'$,
- $\mathcal{I}_{\text{rb}_i}$ est l'ensemble des mesures subjectives associées à $\mathcal{R}'$, à partir de RB_i,
- $\mathcal{D}$-sep et l'ensemble des d-séparations calculées pour chaque règle à partir de RB_i.

La fonction de calcul d'intêrêt subjectif (vis-à-vis du réseau bayésien), ainsi que le calcul de ce que nous appelons les parties « d-séparées » d'une règle, font l'objet de la section 3.5.1.

L'activité liée à cette fonction est représentée dans la figure 3.4.

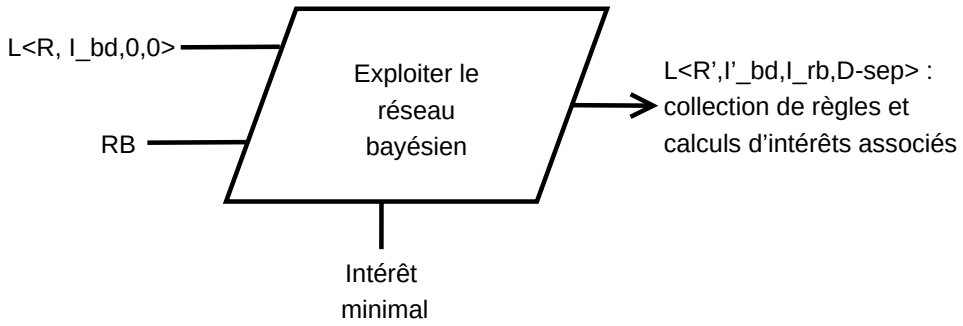

FIG. 3.4 – Activité « Exploiter le réseau bayésien ».

Importance du rôle de l'expert

Comme nous l'avons précisé, l'expert est situé au centre de notre processus. Ainsi nous avons été amenés à développer une interface spécifique pour l'aide à l'analyse des règles d'association. Le but de cette interface est de permettre l'étude d'un volume potentiellement important de règles et de faciliter leur manipulation. L'activité associée est détaillée dans la figure 3.5. Nous pensons en effet que sans interface adaptée, cette phase d'analyse est

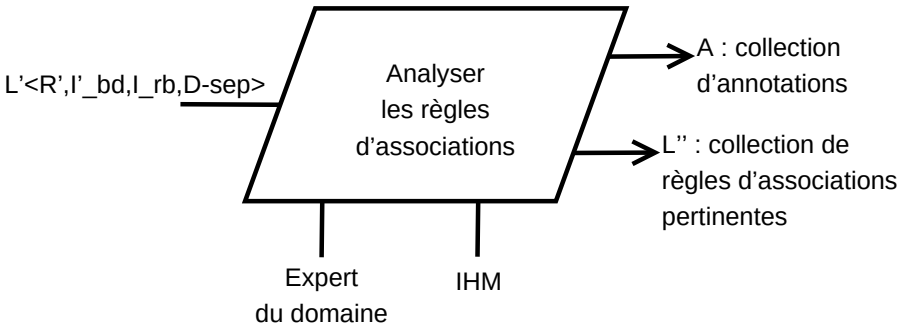

FIG. 3.5 – Activité « Analyser les règles d'association ».

Par le biais de cette interface, l'utilisateur va pouvoir effectuer les différentes itérations du processus de découverte de connaissance. L'étape d'analyse des règles extraites est celle qui requiert le plus d'interactions avec l'expert du domaine. Pour cela nous avons implémenté les fonctionnalités suivantes : tri et seuillage sur les mesures d'intérêts, filtrage syntaxique, tests d'hypothèses, *annotations des règles*, visualisation des parties « d-séparées » et de l'impact des annotations, et *sélection d'un ensemble de règles pertinentes*.

Nous reviendrons sur chacune de ces fonctions, mais on peut retenir que les annotations sont un moyen de mémoriser la présence – dans la collection de règles – de *motifs*[1] déjà connus, non valides, non pertinents, ou intéressants. Ces annotations sont ensuite interprétées *visuellement*, sur l'ensemble des règles étudiées. La deuxième fonction de ces annotations et d'alimenter l'étape qui consiste à faire évoluer le réseau bayésien en fonction des découvertes réalisées par l'expert.

Cette application a été développée en partenariat avec un étudiant de l'Université Paul Sabatier à Toulouse (Mehdi Rabah). Elle est présentée dans l'annexe A.

[1] On utilise ici le terme de *motif* afin de distinguer la règle qui constitue l'annotation et la règle d'association en elle même.

Évolution du modèle des dépendances du domaine

Cette activité (figure 3.6) prend en entrée la collection d'annotation issue de la phase d'analyse des règles, ainsi que la réseau bayésien RB_i.

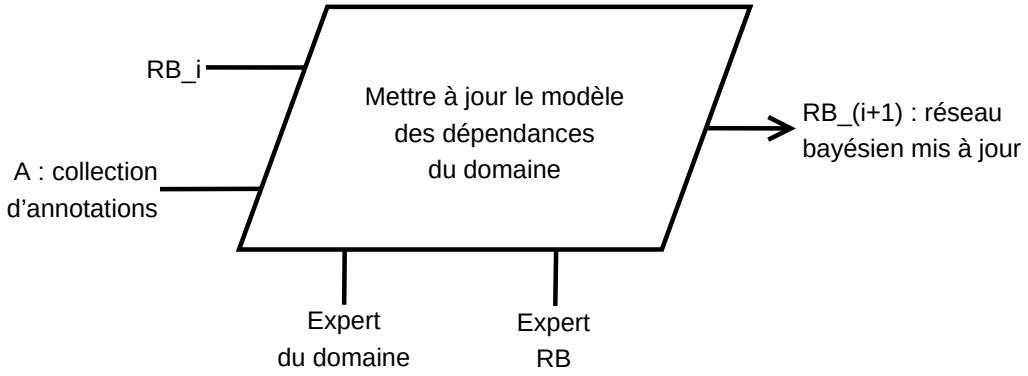

FIG. 3.6 – Activité « Mettre à jour le réseau bayésien ».

L'objectif de cette phase est de décider si l'on va intégrer, ou non, les différentes connaissances portées par les annotations. Ainsi, la définition d'une nouvelle itération de notre modèle va nécessiter la collaboration d'un expert du domaine et d'un expert sur les RB. Pour cela un deuxième composant de notre application a été développé pour permettre les modifications du RB. Il n'y a cependant aucune automatisation dans ce processus : une fois les règles annotées l'expert du domaine enregistre les résultats de son analyse qui sont alors visualisées par l'expert en RB. Comme on va le voir (section 3.6) la syntaxe des annotations a été pensée pour faciliter leur interprétation en vue d'une intégration au modèle (ajout, modification ou suppression d'arcs ou de nœud du graphe). Le cas échéant il faut redéfinir les tables de probabilités impactées par les modifications, il alors possible de s'appuyer sur les données et sur les recommandations de l'expert du domaine via les annotations.

3.2.2 Le processus KARD détaillé

La figure 3.7 décrit l'enchaînement des différentes activités précédemment décrites et présente une vision globale de notre approche pour la découverte de connaissances.

Ce schéma montre la boucle « vertueuse » que l'on a mise en place : au fur et à mesure des itérations le modèle est de plus en plus complet : son utilisation permet de filtrer de plus en plus de motifs non pertinents, réduisant ainsi la collection de règles présentée à l'expert et facilitant par la même le travail. Le postulat pris est que la capacité à découvrir des règles réellement pertinentes augmente en même temps que la proportion de motifs *parasites* – donc perturbateurs – diminue dans la collection de règles étudiées.

Le graphe fait de plus ressortir les deux résultats attendus à l'issue de notre processus de fouille : un réseau bayésien et une collection de règles d'association annotées comme « pertinentes ».

Le RB représente les principales dépendances du domaine et il pourra être réutilisé ultérieurement, par exemple si l'on souhaite effectuer à nouveau le processus de découverte de connaissances sur de nouvelles données. Les dépendances découvertes sur le domaine sont donc capitalisées.

La collection de règles pertinentes présente, de part sa nature, un intérêt pour les experts du domaine. Ainsi la découverte de règles peut permettre d'expliquer un comportement inattendu, ou remettre en cause un fonctionnement jusqu'à alors perçu comme établi, ...

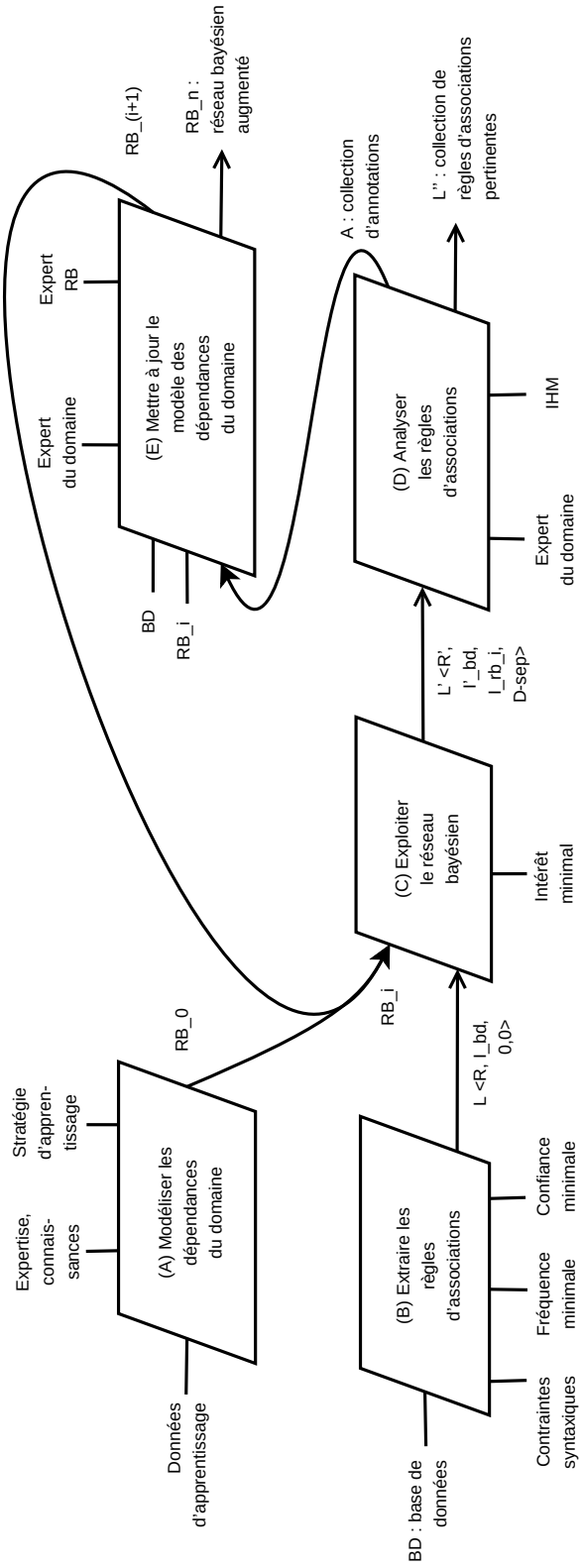

FIG. 3.7 – Proposition de processus de découverte de connaissances

3.3 Le cas d'application « *Visit Asia* »

Afin de présenter les techniques et la méthodologie proposées on considère, dans un premier temps, un cas d'application sur des données simulées. Ce cas d'application va nous permettre, dans la suite de ce chapitre, d'illustrer nos contributions par le biais d'exemples concrets. En dernier lieu il nous permettra de valider de manière expérimentale l'ensemble du cycle de découverte de connaissances proposé.

Pour cela, on s'intéresse au RB de référence, tel qu'il est défini pour le domaine « *Visit Asia* », bien connu de la communauté des Réseaux Bayésiens. Cet exemple a initialement été présenté pour la première fois dans [LS88]. Nous avons décidé de partir de *Visit Asia* car, bien que faisant intervenir un nombre restreint de variables, il présente des similarités avec le cas d'application sur les données d'interruptions opérationnelle que nous étudions au chapitre 4.

Le contexte associé à ce RB est une *simulation* de l'étude de pathologies particulières chez un patient, ainsi que les causes possibles qui favorisent leur apparition. La figure 3.8 ci-dessous nous montre le RB *qui représente le mieux les connaissances relatives à ce domaine*. Il va donc s'agir pour nous du RB *de référence*, on le désignera par RB_ref.

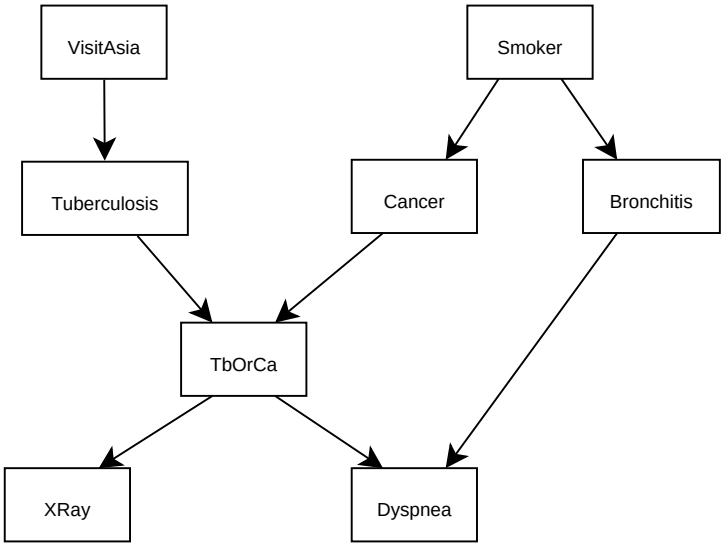

FIG. 3.8 – Réseau bayésien de référence sur le domaine *Visit Asia* (RB_ref).

L'utilisation qui est faite de ce réseau consiste à diagnostiquer une maladie, à partir d'un ensemble de symptômes et de faits *contextuels*, ou réciproquement à déterminer les causes les plus probables de cette maladie. Le réseau *Visit Asia* modélise les faits médicaux suivants :

1. La dyspnée (*dyspnea*) traduit une difficulté à respirer. Elle peut être due à la tuberculose, au cancer du poumon, a une bronchite, ou à aucune de ces maladies.
2. Un voyage récent en Asie augmente les chances de tuberculose.
3. Un patient fumeur aura plus de risques d'être atteint d'un cancer et/ou d'une bronchite.
4. Les résultats d'un examen aux rayons-X ne permettent pas de discriminer entre un patient atteint d'un cancer du poumon ou d'une bronchite.
5. La présence ou l'absence de dyspnée ne permet pas non plus de discriminer ces deux maladies.

Une des spécificités de ce réseau est la variable *VisitAsia* qui nous renseigne sur le fait que le patient a, ou non, effectué un voyage récent en Asie. Comme on peut le voir (figure 3.9), le RB de référence modélise une relation de cause à effet entre le fait d'avoir visité l'Asie et le fait d'être atteint de la tuberculose. Cela se traduit par la présence d'un arc orienté reliant le nœud *VisitAsia* en direction du nœud *Tuberculosis*. Si l'on s'intéresse à la table de probabilité conditionnelle (ou CPT) du nœud *Tuberculosis* on voit qu'elle définit explicitement et quantitativement l'influence de la valeur de *VisitAsia* (*Visit* ou *NoVisit*) sur les valeurs que peut prendre *Tuberculosis*, à savoir : *Present* ou *Absent*).

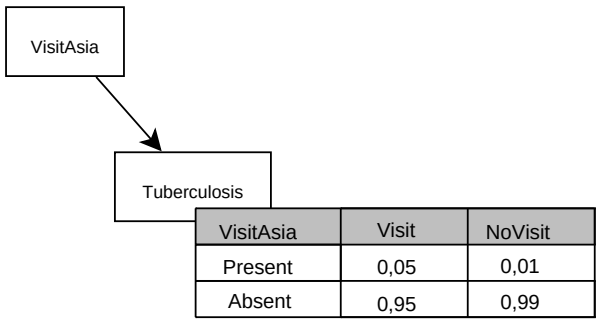

FIG. 3.9 – Exemple de représentation de l'influence d'une variable dans le RB *Visit Asia*.

On peut traduire en langage naturel la relation entre ces deux variables de la façon suivante : « Si le patient a voyagé en Asie, alors la probabilité pour qu'il soit atteint de tuberculose sera plus élevée (i.e., que s'il n'avait pas effectué ce voyage) ».

3.4 Génération d'un ensemble concis de règles d'association à partir des libres, δ-libres

Cette section présente une étude de différentes collection de règles non redondantes. On va pour cela étudier les propriétés de représentations qui permettent de

générer un ensemble de règles valides, mais dont la confiance est approximative (i.e., elle ne peut être calculée avec certitude). L'objectif de cette section est d'apporter une meilleure compréhension des propriétés de ces ensembles de règles. Nous proposons aussi quelques pistes pour l'extraction de collections de règles d'association δ-approximatives et δ-générale, notamment lorsqu'une approximation de la fréquence de la partie droite de la règle – et donc indirectement de la mesure de confiance – n'est pas rédhibitoire pour l'utilisateur.

Contexte, itemsets δ-libres, δ-fermeture

[Pas00] a présenté une collection de règles non-redondantes générées à partir des itemsets libres et des fermés. Cependant, dans certains cas (données fortement corrélées, nécessité d'extraire à des seuils de fréquence très faible) il n'est pas possible de générer l'ensemble des itemsets libres. Ainsi, nous avons choisi d'étudier d'autres types de représentation pour la génération de règles non redondantes, offrant plus de souplesse vis-à-vis de la contrainte de fréquence. Pour cela nous allons commencer par introduire rapidement les principales notions relatives à l'extraction de règles non redondantes.

Considérons la base de données suivante[2] :

T_{id}	t_i.item
1	{A,C,D}
2	{B,C,E}
3	{A,B,C,E}
4	{B,E}
5	{A,B,C,E}
6	{B,C,E}

TAB. 3.1 – Exemple de base de données.

Des règles d'association utilisant les itemsets δ-libres ont été introduites pour la première fois dans [BBR00], il s'agit des règles dites δ-fortes. Le paramètre δ est censé être petit par rapport au seuil de fréquence utilisé pour l'extraction, ce qui assure des règles à forte confiance (i.e., peu d'exceptions sur la partie droite). La définition d'une règle δ-forte est la suivante :

Définition 3.1 (Règle δ-forte) *Étant donné une base de données bd définie sur Items, un seuil de fréquence minimale γ, et un entier $\delta > 0$, une règle δ-forte sur bd*

[2]Il s'agit du même exemple que celui présenté dans [Pas00], il nous permettra de comparer les différentes collections de règles.

est une règle d'association $X \rightarrow Y$, *et* $Freq(X \cup Y) \geq \gamma,\ Freq(X) - Freq(X \cup Y) \leq \delta,\ X \cup Y \subseteq Items, avec\ Y \neq \emptyset$.

Exemple. Une règle δ-forte accepte donc, au plus, δ exceptions sur sa partie droite. Dans bd un exemple de règle « 1-forte » est AC → E, par contre BE → A n'est pas une règle 1-forte.

Un sous-ensemble des règles δ-fortes sont les règles δ-fortes dont la partie droite est composée d'un seul item. Ces règles sont particulièrement intéressantes pour des problématiques de classification [CB02] car il est possible, à partir de critères sur δ et sur le seuil de fréquence minimale d'obtenir un ensemble de règles au corps minimal caractérisant des classes (i.e., la partie droite des règles est une classe).

L'ensemble des règles δ-fortes peut être construit à partir de l'ensemble des itemsets δ-libres, qui forment le corps de la règle. On rappelle la définition d'un itemset δ-libre.

Définition 3.2 (Itemset δ-libre) *Soit S un itemset libre. S est δ-libre s'il n'existe pas de règle δ-forte $X \rightarrow Y$ telle que $X \cup Y \subseteq S,\ Y \neq \emptyset$.*

Exemple. Sur la base de données bd représentée dans la table 3.1 on extrait l'ensemble des itemsets 2-fréquents libres (i.e., $minfreq = 2,\ \delta = 0$) : $\{A, B, C, E, AB, AE, BC, CE\}$. L'ensemble des itemsets 2-fréquents 1-libre est quant à lui composé uniquement de $\{A\}$. En effet, on peut voir, par exemple, que AE n'est pas un itemset 1-libre car la règle A → E admet 1 exception, de même B n'est pas 1-libre car il admet 1 exception sur bd, etc.

Le concept de δ-libre est intimement lié à la notion de δ-fermeture (ou *quasi*-fermeture) introduite par [BBR00].

Définition 3.3 (δ-fermeture) *Soit S un itemset $\subseteq Items$ et δ un entier positif. La δ-fermeture de S, est le plus grand sur ensemble de S définit comme suit :*

$$ferm_\delta(S) = \{I \in Items \mid Freq(S) - Freq(S \cup \{I\}) \leq \delta\}.$$

Exemple. Si l'on poursuit notre exemple sur les données de la table 3.1, on peut calculer la 1-fermeture de A qui est donnée par : $ferm_1(A) = \{A, B(-1), C, E(-1)\}$. Pour chaque item X_i de la δ-fermeture on note entre parenthèses la différence entre $Freq(A)$ et $Freq(A \cup \{X_i\})$ (e.g., ($Freq(A) = 3$, et $Freq(AB) = 2$, d'où $s_B = 1$).

Ainsi l'opérateur de δ-fermeture permet d'extraire un ensemble d'itemsets dont le support est borné. Si on note respectivement s_B, s_C, s_E le nombre « d'exceptions » des itemsets qui composent la δ-fermeture de A, alors on peut en déduire que :

$$Freq(A) - max(s_B, s_C, s_E) \leq Freq(ferm_1(A)) \leq Freq(A) - s_B - s_C - s_E.$$

Plus généralement on pourra dire qu'étant donné un itemset S, et $X = \{X_1, X_2, \ldots, X_n\}$ sa δ-fermeture, alors la fréquence de X sera :

$$Freq(S) - max(s_{X_i}) \leq Freq(X) \leq Freq(S) - \sum_{i=1}^{n} s_{X_i}.$$

On vérifiera facilement que $Freq(S) - max(s_{X_i}) \leq Freq(S) - \sum_{i=1}^{n} s_{X_i}$. De plus, on en déduit que pour obtenir des résultats « cohérents » il faut que δ – pour tout $ferm_\delta(S) = X \neq \emptyset$ – soit au moins strictement inférieur à $Freq(S)/card(X)$. En effet, dans le cas contraire cela signifierait que la δ-fermeture calculée aurait éventuellement une fréquence nulle ou négative (i.e., il ne serait pas supporté par les données de bd). Ainsi dans notre exemple précédent la fréquence de la 1-fermeture de A peut être égale à 0 étant donné que $Freq(A) = 3,\ \delta = 1$ et $card(BCE) = 3$ (en pratique ce n'est pas le cas).

Étude des règles d'association formées à partir des itemsets libres et de la δ-fermeture

L'utilité de la δ-fermeture a été démontrée, dans le cadre de l'extraction de représentations condensées d'itemsets fréquents, lorsque l'on travaille sur des données fortement corrélées a des seuils de fréquence relativement bas [BBR00].

Maintenant, quand est-il d'une collection de règles utilisant les itemsets δ-fermés ? Soit $minfreq$ un seuil de fréquence minimum, et δ un entier positif tels que $\delta < minfreq/card(Items)$. On envisage alors l'ensemble de règles d'association tel que :

$$\{R : \mathrm{X} \rightarrow \mathrm{Y} \mid X \subseteq Items,\ Y = ferm_\delta(X) \backslash X,\ Y \neq \emptyset\}.$$

Il s'agit donc d'un ensemble de règles d'association dont la confiance est, a priori, bornée de la façon suivante :

$$1 - \delta \times card(Y) \leq conf(R : \mathrm{X} \rightarrow \mathrm{Y}) \leq 1 - max(s_{Y_i}).$$

Exemple. La table 3.2 montre l'ensemble des règles d'association calculées à partir des 1-fermetures de tous les itemsets 2-fréquents.

Les valeurs de confiance indiquées entre parenthèses sont « comptées » sur les données. En effet, comme nous l'avons vu, la δ-fermeture ne permet de donner qu'une approximation bornée de la valeur réelle. Si on étudie maintenant les règles présentées dans le tableau 3.2 on peut constater que certaines règles sont redondantes entre elles :

Num	Règle R : X → Y	$Freq(X)$	$Conf(R)$	$R \in \delta$-approx/δ-gen ?
1	B → C(-1)E	5	(0,8)	oui / non
2	C → B(-1)E(-1)	5	(0,8)	oui / non
3	E → BC(-1)	5	(0,8)	oui / non
4	BE → C(-1)	5	(0,8)	non / non
5	BC → E	4	1,0	oui / non
6	CE → B	4	1,0	oui / non
7	A → B(-1)CE(-1)	3	(0,66)	oui / oui
8	AC → B(-1)E(-1)	3	(0,66)	non / non
9	AB → CE	2	1,0	oui / non
10	AE → BC	2	1,0	oui / non
11	ABC → E	2	1,0	non / non
12	ABE → C	2	1,0	non / non
13	ACE → B	2	1,0	non / non

TAB. 3.2 – Règles extraites sur bd pour $minfreq = 2$ et $\delta = 1$.

- La règle n°4 peut être générées à partir des règles 1 et 3.
- Les règles 7 et 8 sont redondantes entre elles, la règle 7 est cependant plus générale car son corps est minimal par rapport à 8, c'est donc celle là qu'on souhaite conserver.
- Les règles 11, 12, 13 peuvent être générées à partir des règles 9 et 10.

Pour éliminer la redondance il nous faut donc sélectionner les règles qui, pour une fréquence donnée, ont un corps minimal. En fait l'ensemble des règles non-redondantes de la collection issue de la table 3.2 est *exactement* l'ensemble des règles dont le corps est un itemset 2-fréquents libres. Ce résultat se démontre facilement en exploitant les propriétés des itemsets libres, qui sont les itemsets minimaux des classes d'équivalence. Cela nous amène naturellement à la définition d'une première base génératrice de règles d'association.

Définition 3.4 (Base δ-approximative de règles d'association) *La base δ-approximative de règles d'association de la façon suivante :*

$$\delta\text{-}approx = \{R : X \rightarrow ferm_\delta(X) \backslash X \mid X \in Libre\}.$$

Ainsi, dans notre exemple, la base δ-approximative de règles d'association, est constituée par l'ensemble des règles $\{1, 2, 3, 5, 6, 7, 9, 10\}$.

Dans le cas où $\delta = 0$ la base calculée est exactement la base des règles d'association exactes (i.e. de confiance égale à 1), on retrouve alors le résultat présenté par [Pas00] sur la même base de données.

Cette base est intéressante car elle offre un bon compromis entre compacité et précision d'extraction, cependant elle nécessite le calcul des itemsets libres, ce qui n'est pas toujours possible selon que les données sont, ou non, fortement corrélées, et que l'on souhaite extraire des règles avec une fréquence relativement faible.

Étude des règles d'association générées à partir des itemsets δ-libres et de la δ-fermeture

Il est possible de généraliser encore plus la base précédente en utilisant cette fois les itemsets δ-libres.

Si l'on regarde les résultats précédents, on constate que les règles 5 et 6 sont redondantes par rapport à 1 et 2, à *l'erreur de confiance près*. De même pour les règles 9 et 10 par rapport à la règle 7. Enfin, les règle 1, 2, 3 sont moins générales que la règle 8 (toujours à l'erreur de confiance près), qui elle même est moins générale que la règle 7.

En conclusion, si l'on tolère une erreur sur la confiance des règles générées, on peut représenter l'ensemble des règles de la table 3.2 uniquement par la règle n°7. Le corps de la règle 7 est en fait le seul itemset 2-fréquent 1-libre calculé sur bd. Cela nous permet de définir la base des règles approximatives les plus générales sur bd :

$$\delta\text{-gen} = \{R : X \rightarrow ferm_\delta(X) \backslash X \mid X \in Libre_\delta\}.$$

Exemple. Comme on vient de le voir une seule règle constitue cette base il s'agit de la règle : R : A $\rightarrow$ B(-1)CE(-1). Nous allons maintenant étudier la possibilité de générer les règles de δ-approx à partir de cette règle, et nous allons voir les erreurs d'estimation de fréquence apportées par cette représentation.

On sait, grâce à la δ-fermeture que $Freq(AB) = Freq(A) - 1 = 2$, on peut donc estimer la confiance de la règle AB $\rightarrow$ CE qui sera d'au moins $\frac{Freq(ABCE)}{Freq(AB)}$, soit $Conf(\text{AB} \rightarrow \text{CE} \geq = 0,5$. En réalité la confiance de cette règle est égale à 1. Dans ce cas, on a donc une erreur d'estimation, sur la borne inférieure, de l'ordre de 50%.

Ce résultat n'est pas étonnant, car dans notre exemple δ ne respecte pas les critères que nous avons défini ; il ne faut donc pas perdre de vue qu'il s'agit là d'un exemple didactique qui vise à montrer qu'il est *possible* d'estimer la fréquence des différentes règles à partir de l'ensemble des itemsets δ-libre et de leur δ-fermeture. En pratique, lorsque δ est plus faible par rapport au seuil de fréquence utilisé, l'erreur d'estimation est elle aussi plus faible.

Autre exemple de calcul, on sait que $Freq(BC) \geq Freq(A) - 1 = 2$, et que $Freq(BCE) \geq 1$, on peut donc estimer $Conf(\text{BC} \rightarrow \text{E}) \geq 0,5$.

On a vu qu'il était possible de générer l'ensemble des règles δ-approx à partir des règles de δ-gen. On utilise pour cela une approximation sur les fréquences des itemsets qui composent ces règles, à partir des informations de la δ-fermeture, et de la fréquence des itemsets δ-libres. En pratique, on veillera à ce que δ soit très inférieur à $minfreq/card(Items)$ pour assurer une erreur d'approximation de la fréquence réduite.

Discussion

Nous avons présenté deux représentations de règles d'association paramétrées par un entier positif δ. Il s'agit de bases qui permettent de générer un ensemble de règles d'association valides. Lorsque $\delta = 0$ ces bases sont équivalentes aux résultats présentés par l'algorithme CLOSE, proposé par N. Pasquier et al. Lorsque $\delta > 0$ on constate que les règles extraites sont plus générales, le volume de règles présentées à l'utilisateur et donc plus réduit. En contrepartie, la précision des règles obtenues est incertaine, d'une part la confiance est bornée par δ, d'autre part, pour la collection utilisant les δ-libres la fréquence du corps des règles que l'on peut générer est elle aussi incertaine.

Ces bases sont intéressantes car elles contiennent des règles ayant ont un corps minimal, pour une fréquence donnée, et une tête maximale : ce sont donc les règles les plus *informatives* pour l'utilisateur. De plus le programme AC-Like (qui est une implémentation de l'algorithme MIN-EX [BBR00]) permet d'extraire l'ensemble des itemsets δ-libres ainsi que leur δ-fermeture. Cet algorithme s'est révélé très efficace, en termes de temps d'extraction, sur les données de notre cas d'application industriel. Il nous a effectivement permis de fixer des seuils de fréquence relativement bas, tout en conservant une collection de règles extrêmement compacte.

3.5 Exploitation d'un Réseau Bayésien pour la découverte de règles d'associations pertinentes

3.5.1 Définition d'une mesure de pertinence des règles, vis-à-vis d'un réseau bayésien

Concernant les règles d'association, nous savons qu'il peut être très tentant de vouloir les interpréter comme une formulation de la causalité entre deux ensembles de variables. Mais pour déterminer si cette notion de causalité existe réellement il nous manque des informations contextuelles que les données seules ne peuvent pas fournir.

La première utilisation du RB est l'inférence, qui consiste à calculer des probabilités conditionnelles d'événements reliés les uns aux autres par des relations de cause à effet. Un réseau bayésien modélisé avec soin permet de décrire – et donc de mesurer

– de manière quantitative le lien de causalité pouvant exister entre deux variables.

Ainsi on voit qu'il peut être intéressant de comparer la mesure de « causalité » estimée par le biais de la confiance d'une règle d'association avec celle qui est modélisée dans le RB.

Soit bd une base de données binaires de schéma $\langle T_{id}, Items \rangle$, tel que $Items = \{A_1, A_2, \ldots, A_n\}$.

Soit RB un réseau bayésien défini par un ensemble de noeuds correspondants aux attributs de $Items$ et par $E \subset Items \times Items$ l'ensemble des arcs du graphe. A chaque nœud on associe une distribution de probabilité conditionnelle $P(A_i|\mathrm{Par}(A_i))$, où $\mathrm{Par}(A_i) = \{A_j|(V(A_j), V(A_i)) \in E\}$ représente les parents du nœud A_i. Pour une discussion détaillée sur les réseaux bayésiens le lecteur pourra consulter [Pea88, NWL+04].

Soit $R : \mathrm{X} \rightarrow \mathrm{Y}$ une règle d'association où X et Y sont des itemsets tels que $X, Y \subseteq Items$, $Y \neq \emptyset$ et $X \cap Y = \emptyset$. La fréquence d'un itemset X dans bd, notée $p_{bd}(X)$, est l'ensemble des lignes de bd qui contiennent X par rapport à la taille de bd. Cette fréquence dénote, sous réserve que la taille de bd soit suffisamment grande, la probabilité que tous les items X_i de l'itemset $X \subseteq Items$ soient observés dans les données de bd (i.e., ils prennent la valeur « vrai »).

De même on définit $p_{bd}(R : \mathrm{X} \rightarrow \mathrm{Y}) = p_{bd}(X \cup Y)$ la probabilité qu'une règle d'association soit observée sur les données. Il s'en suit que la confiance d'une règle R exprimée en termes de probabilités s'écrit de la façon suivante :

$$\mathrm{conf}_{bd}(R : \mathrm{X} \rightarrow \mathrm{Y}) = \frac{p_{bd}(R)}{p_{bd}(X)}.$$

Ainsi, étant donné une base de donnée bd et un réseau bayésien RB, nous avons défini une mesure subjective de l'intérêt d'une règle d'association vis-à-vis d'un réseau bayésien. Cette mesure est inspirée de [JS04]. Elle se base sur le calcul de la différence entre la confiance de la règle estimée à partir des données et la probabilité inférée par le réseau bayésien d'observer les attributs de la partie droite de cette règle, sachant que l'on observe les attributs de la partie gauche.

Pour une règle d'association $R : \mathrm{X} \rightarrow \mathrm{Y}$ cette mesure d'intérêt s'écrit :

$$\begin{aligned} \mathrm{Int}_{rb}(R) &= |\mathrm{conf}_{bd}(R) - \mathrm{conf}_{rb}(R)| \\ \text{où } \mathrm{conf}_{bd}(R) &= \frac{p_{bd}(X \cup Y)}{p_{bd}(X)} \\ \text{et } \mathrm{conf}_{rb}(R) &= \prod_{i=1}^{m} p(Y_i|X_1, \ldots, X_i, \ldots, X_n) \end{aligned}$$

Exemple. Pour illustrer le calcul de cette formule prenons l'exemple de *Visit Asia* et son réseau RB_ref tel qu'il est présenté dans la figure 3.8, page 74. A partir de données disponibles sur le domaine nous extrayons, à titre d'exemple, les règles d'association de la table 3.3.

Num	Règle d'association	conf_{bd}	conf_{rb_ref}
1	Dyspnea VisitAsia → XRay	0,98	0,86
2	Bronchitis Cancer Dyspnea Smoking → TbOrCa XRay	0,97	0,07

TAB. 3.3 – Exemples de règles d'association extraites sur *Visit Asia* à partir de RB_ref.

Commençons par expliciter le calcul de Int_{rb} sur la règle n°1 :

$$\text{conf}_{rb\_ref} = p(Xray = Abnormal | Dyspnea = Present, VisitAsia = Visit).$$

Les calculs d'inférence bayésienne[3] nous donnent alors : $\text{conf}_{RB\_ref} = 0,24$. La confiance calculée à partir des données étant de $\text{conf}_{\text{bd}} = 0,98$, on a $\text{Int}_{rb} = |0,98 - 0,24| = 0,74$. La règle est donc jugée valide et potentiellement intéressante puisqu'elle représente un événement qui est statistiquement relativement rare.

Imaginons maintenant que l'on extraie la même règle mais cette fois les nœuds *VisitAsia* et *XRay* sont complètement déconnectés du reste du graphe (les tables de probabilités conditionnelles sont évidemment redéfinies en conséquence). Le même calcul d'intérêt vis-à-vis de ce RB consisterait alors à déterminer $p(XRay = Abnormal)$ soit $0,11$, ce qui nous donnerait $\text{Int}_{rb} = 0,87$. Dans ce cas l'intérêt par rapport au RB est plus élevé puisque l'association exprimée par la règle ne se retrouve pas dans le modèle. Cela correspond parfaitement au comportement attendu de notre mesure.

Étudions à présent la règle n°2. La formule nous donne :

$$\begin{aligned}
\text{conf}_{rb_ref} &= p(TbOrCa = True | Bronchitis = Present, Cancer = Present, \\
&\quad Dyspnea = Present, Smoking = Smoker) \\
&\times \; p(XRay = Abnormal | Bronchitis = Present, Cancer = Present, \\
&\quad Dyspnea = Present, Smoking = Smoker) \\
&= 1,00 \times \; 0,98 \\
&= 0,98. \\
\text{Int}_{rb_ref} &= |0,97 - 0,98| \\
&= 0,01.
\end{aligned}$$

[3]Tous les calculs d'inférence peuvent être reproduits en utilisant le logiciel libre *Bayesian Network Tools in Java* disponible à l'adresse suivante : `http://sourceforge.net/projects/bnj`. Le réseau *Visit Asia*est lui aussi disponible via l'application.

3.5.2 Extraction des parties d-séparées, dépendances principales

Il est intéressant de pouvoir exploiter les composantes graphiques du RB dans le cadre de l'analyse de règles d'association. Un arc orienté reliant deux variables X et Y d'un RB, et une règle association X $\rightarrow$ Y ont une représentation graphique très proche. Dans les deux cas une notation d'orientation intervient. Intuitivement on pourrait traduire cette orientation par la phrase : « Le fait d'observer X m'apporte une connaissance supplémentaire sur Y ». Il y a donc un flot d'information qui circule entre ces deux variables. Ainsi, à partir d'un RB et d'une collection extraites sur le domaine du RB (i.e. les variables du RB sont les mêmes que les variables utilisées par les règles d'association) on peut se poser la question de savoir si ce flot d'information est représenté de manière identique dans le RB et dans les règles d'associations ? Quelles sont précisément les différences que l'on peut constater ? Ces différences permettront alors de d'identifier si les règles sont non-valides (i.e., l'information quelles portent est contraire à la définition des dépendances du RB) ou au contraire si elles peuvent être potentiellement pertinentes (i.e., découverte d'une règle qui montre une association valide mais non prise en compte dans le RB).

Comme nous l'avons évoqué la circulation de l'information dans le RB obéit à la propriété de d-séparation (section 2.5.2, page 47). Pour rappel le test de d-séparation entre X et Y conditionnellement à Z (où X, Y et Z sont des sous-ensembles disjoints de l'ensemble des nœuds associés au graphe du RB) s'écrit : $\langle X|Z|Y\rangle$. X est d-séparé de Y par Z signifie alors que la présence de Z *bloque* le cheminement de l'information de X vers Y. Formulé différemment, on dira qu'en présence de Z le fait d'observer X n'apporte pas de connaissances supplémentaires sur Y.

Ainsi pour déterminer les différences en termes de circulation de l'information entre les règles et le RB, nous allons appliquer la propriété de d-séparation sur la collection de règles extraites, par rapport à la structure du RB. Pour chaque règle d'association R : X $\rightarrow$ Y, on calcule le test de d-séparation $\langle X_i|X\backslash X_i|Y_j\rangle$ pour tous les $X_i \in X$ et pour tous les $Y_j \in Y$ (i.e., X_i et Y_j sont des items de la règle).

Si X est de taille 1 alors Z est égal à $\emptyset$ et aucun Y_j n'est d-séparé de X.

Dans le cas où le nombre d'items de X est strictement supérieur à 1, on étudie la matrice booléenne qui contient tous les résultats des tests de d-séparation entre les X_i et les Y_j. Si pour un item X_i ou Y_j donné tous les résultats de d-séparation sont positifs alors cet item est ajouté à l'ensemble que nous appelons : « partie d-séparée » de la règle, ou $\mathcal{D}$-sep. Concrètement cela signifie qu'une association a été trouvée dans les données, mais qu'elle n'est pas modélisée dans le RB. Son ensemble complémentaire est appelé *ensemble des dépendances principales*, on le note $\mathcal{D}$-core.

Exemple. Soit la règle d'association R : ABC $\rightarrow$ D. Notre algorithme calcule les tests de d-séparation suivants, et uniquement ceux là :

- (1) $\langle A|BC|D\rangle$?
- (2) $\langle B|AC|D\rangle$?
- (3) $\langle C|AB|D\rangle$?

Si (1) est vrai, et (2) et (3) sont faux alors : $\mathcal{D}$-sep$(R) = \{A\}$. On peut traduire cela par : « sachant qu'on observe B et C, A n'apporte rien de plus sur la connaissance de C ». L'expert doit alors étudier la règle R pour savoir si l'itemset A présente ou non un intérêt.

Par contre si (1), (2) et (3) sont vrais alors on a $\mathcal{D}$-sep$(R) = \{ABCD\}$. On traduira ici : « Les variables A, B et C sont indépendantes entre elles ». Là aussi, le fait de mettre en avant la présence d'une règle dont la partie d-séparée est non vide, va inciter l'expert du domaine à intervenir pour valider ou infirmer la pertinence de la règle en question (i.e. par le biais des annotations). Puisqu'il y a une différence entre l'association constatée à partir de données et la modélisation de cette association dans le RB, alors l'expert doit déterminer si la règle d'association est fortuite ou si au contraire une découverte pertinentes à été mise en avant.

Cependant, on remarquera que l'algorithme que nous proposons donne seulement une approximation de la décomposition de la règle en parties d-séparées / dépendances principales. En effet toutes les combinaisons d'itemsets (i.e. les sous-ensembles de X et de Y) ne sont pas testées pour la d-séparation.

Reprenons l'exemple précédent, si nous avions effectué les calculs de d-séparation sur tous les sous-ensembles d'itemsets, nous aurions eu en plus à effectuer les tests suivants :

- (4) $\langle AB|C|D\rangle$?
- (5) $\langle AC|B|D\rangle$?
- (6) $\langle BC|A|D\rangle$?

Suivant les résultats de ces tests la composition des parties d-séparées de la règle peut être différente. Ainsi, dans le cas où les tests (1), (2) sont vrais mais que (4) est faux, on peut donner l'interprétation suivante : « Sachant que j'observe C, le fait d'observer A et B simultanément, m'apporte une connaissance supplémentaire sur C ». Dans ce cas $\mathcal{D}$-sep $= \emptyset$.

3.6 Le rôle de l'expert dans le processus de découverte

Nous disposons d'un algorithme capable d'extraire une collection concise de règles d'association à partir de grands volumes de données, d'un formalisme pour modéliser certaines connaissances a priori de l'expert, ainsi que d'une mesure prenant en compte ces connaissances pour évaluer l'intérêt des règles d'associations générées. Tous ces outils sont mis au service de l'expert en charge de l'analyse des règles. Ils vont lui per-

mettre d'acquérir une compréhension plus fine des règles manipulées et implicitement une meilleure compréhension du domaine.

L'expert doit être en mesure de transférer sa « compréhension » des règles vers le modèle des dépendances utilisé. De cette façon on pense découvrir des règles de plus en plus pertinentes. Cette section décrit le rôle que joue l'expert dans notre approche ainsi que le système d'annotations proposé.

3.6.1 Nécessité des annotations

La tâche de l'expert est d'analyser et interpréter les règles extraites. Pour cela on met à sa disposition un système d'annotation qui va lui permettre de porter un jugement sur les règles, et en particulier sur les motifs qui la compose. Par « motif » d'une règle d'association $R : \mathrm{X} \rightarrow \mathrm{Y}$ on entend ici une « sous règle » d'association $M : X' \rightarrow Y_i$ telle que $X' \subseteq X$ et $Y_i \in (X \backslash X') \cup Y$.

En pratique il n'est pas rare de retrouver un motif déjà connu ou inintéressant dans un grand nombre de règles d'association. L'expert sait reconnaître ces motifs et il nous est apparu nécessaire de lui donner les moyens de les indiquer par le biais d'une syntaxe bien définie. Ces annotations vont avoir un intérêt double : d'une part le moteur d'affichage des règles va pouvoir rendre compte de la nature des différents motifs qui composent chaque règle ; d'autre part ces annotations sont utilisées pour faciliter la mise à jour du modèle.

3.6.2 Différents types d'annotation

Un des problèmes pour rendre possible l'étape d'annotation des règles est la définition d'une syntaxe rapidement assimilable, et permettant de décrire les différentes informations apportées par une règles d'association. Pour cela nous avons spécifié une notation sous la forme d'une grammaire BNF telle que présentée dans la figure 3.10.

```
liste-annotation  ::=  liste-annotation annotation | annotation
annotation        ::=  '('liste-élément'=>' élément ';' categorie')'
liste-élément     ::=  liste-élément 'et' élément | élément |
élément           ::=  attribut | attribut'='valeur
categorie         ::=  'K:'probabilité | 'NV' | 'NP' | 'I'
```

FIG. 3.10 – Grammaire BNF pour l'annotation des règles d'association.

Ainsi, *si l'on ne considère que les attributs* – et non pas les couples (attribut, valeur) –, toute annotation d'une règle d'association $R : \mathrm{X} \rightarrow \mathrm{Y}$ se présente sous la forme d'une sous partie $A : \mathrm{X'} \rightarrow \mathrm{Y'}$ de R telle que $X' \subseteq X$, $Y' \subseteq Y$ et $\mathrm{card}(Y') = 1$.

Le but de cette notation est de permettre à l'expert de distinguer la nature des dépendances exprimées dans les règles, dans le cadre de la découverte de règles d'association pertinentes. Les dépendances – et donc les annotations – peuvent être de quatre types :

(K) La règle contient une association connue de l'expert, mais non prise en compte par la modélisation actuelle du réseau bayésien. Il est alors possible de modifier la structure du RB afin d'intégrer la notion de causalité à l'origine de ce motif. A l'itération suivante du processus, l'utilisateur ne verra plus apparaître ce type de règles car elles seront jugées inintéressantes.

(NV) L'annotation présente une information fortuite, elle décrit en fait la coïncidence statistique de certains attributs mais l'expert peut affirmer qu'elle n'a pas de valeur en tant que « nouvelle connaissance ».

(NP) Ces annotations décrivent une relation valide mais non pertinente par rapport au contexte dans lequel se situe l'expert.

(I) L'annotation est intéressante. C'est à dire qu'elle « surprend » l'expert du domaine et va demander une analyse approfondie (e.g., une nouvelle itération du processus, voire un retour sur la collecte et la préparation des données, par exemple en introduisant une nouvelle variable).

Exemple. Afin d'illustrer ces différentes catégories d'annotations considérons les règles 1 à 4 du tableau 3.4.

Num	Règle d'association
1	Smoking XRay → Bronchitis
2	Dyspnea Tuberculosis → TbOrCa
3	VisitAsia → TbOrCa Tuberculosis
4	Smoking Bronchitis → VisitAsia

TAB. 3.4 – Exemples de règles d'association fictives sur *Visit Asia*.

Si l'on étudie ces règles, au regard des faits précédemment évoqués sur le domaine *Visit Asia* (voir page 61), on peut formuler les remarques suivantes :

- Un examen aux rayons-X révélant des résultats anormaux ne permet pas de conclure avec certitude sur le fait que le patient ait une bronchite (fait n°4) ou non. Il s'agit donc d'un motif *non valide*.
- On *sait* (fait n°3) qu'un patient fumeur aura plus de risques d'être atteint de bronchite.
- Si on observe un patient atteint d'une bronchite alors on *sait avec certitude* que la variable *TbOrCa* sera instanciée dans les données.
- Par contre il est *intéressant* de s'apercevoir qu'un certain nombre de patients qui déclarent avoir récemment effectué un voyage en Asie montrent les symptômes liés à la tuberculose (fait n°2).

– Enfin on juge comme *non pertinent* de constater qu'une certaine population de patients fumeurs aient voyagé en Asie récemment.

Ces remarques se traduisent par un ensemble d'annotations résumées par le tableau 3.5.

Num	Annotations de l'expert	Règles impactées
1	(XRay → Bronchitis ; NV)	{1}
3	(Smoking → Bronchitis ; K :'assez probable')	{1}
2	(Tuberculosis → TbOrCa ; K :'certain')	{2, 3}
4	(VisitAsia → TbOrCa Tuberculosis ; I)	{3}
5	(Smoking → VisitAsia ; NP)	{4}

TAB. 3.5 – Annotations collectées sur les règles *Visit Asia*. FIXME : compléter

Lorsqu'il rédige ces annotations, l'expert a pour objectif qu'un maximum de règles contenant uniquement des motifs de type (K), (NP), et (NV) soient filtrés lors de la prochaine itération du processus. Ce filtrage peut intervenir par le biais de la mesure d'intérêt subjective (une modification du RB en fonction des annotations collectées doit diminuer l'intérêt de ces règles) ou de manière graphique (impact visuel des annotations). En effet, nous souhaitons que les règles affichées intègrent un maximum d'informations relatives aux annotations, dans le but de faciliter les étapes ultérieures d'analyse.

L'idée est d'affiner progressivement le modèle de connaissance utilisé pour le filtrage des règles d'association en y intégrant les dépendances récemment découvertes. Cependant, l'interprétation des motifs extraits et le choix des modifications a apporter au RB ne sont pas des tâches faciles à réaliser.

3.6.3 Prise en compte des annotations et mise à jour du réseau bayésien

Considérons d'abord le cas des annotations de type (K). A partir de ces annotations, on doit mettre à jour la structure et les paramètres du réseau bayésien. Pour cette étape, il faut répondre principalement à deux types de problèmes.

La première catégorie de problème est relative à la traduction d'une annotation en éléments de modifications du réseau bayésien. La syntaxe que nous avons proposée facilite ce passage. Soit les variables X, Y et Z. Une annotation « (X et $Y \rightarrow Z; C : p$) » sera prise en compte par la création d'un arc de X vers Z et d'un autre de Y vers Z. La table des probabilités est alors modifiée en conséquence, en fixant $p(Z|X,Y) = p$. On procédera de la même façon pour traiter les associations simples de type X → Y.

Le second problème est lié à la modification du réseau bayésien. La définition

des tables de probabilités est un problème délicat et coûteux lorsque les variables manipulées ont un nombre élevé de valeurs possibles. Or, dans un cas d'application réel certaines variables peuvent prendre une centaine de valeurs possibles, rendant toute modification manuelle de la structure du réseau relative à ces variables, extrêmement coûteuse et délicate.

La solution proposée consiste à effectuer un apprentissage automatique des tables de probabilités conjointes puis à soumettre le résultat obtenu à l'expert pour validation. Il pourrait être intéressant de mesurer quantitativement le temps nécessaire pour effectuer ce type d'opérations, et réfléchir sur les modalités d'interactions avec l'expert qui permettraient de faciliter et d'accélérer cette étape.

La deuxième catégorie d'annotations (motifs jugés non valides – NV) est prise en compte indépendamment du réseau bayésien. Chaque motif classé comme « non valide » est ajouté à une base de règles, dont la construction ne sera pas présentée ici. Cet ensemble de règles peut alors servir de filtre pour le post-traitement des règles d'assocations. Si l'expert juge le motif $X \rightarrow Y$ comme étant non valide, il peut décider de masquer ce type d'association en appliquant un filtre sur la collection de règles d'association extraites. Soit le motif $X \rightarrow Y$ jugé non valide par l'expert et une règles d'association $AX \rightarrow BY$ contenant ce motif. Après application du filtre, la règle apparaît sous la forme $A \rightarrow B$.

Enfin, considérons les annotations de type (NP), les motifs non pertinents ne vont pas, par définition, nécessiter une modification de la structure du réseau. Par contre, on leur associe un code couleur afin que l'expert puisse facilement repérer les motifs qu'il a jugé comme non pertinents, au sein de la collection de règles.

3.7 Validation expérimentale de l'approche KARD sur le domaine *Visit Asia*

3.7.1 Objectifs de notre démarche expérimentale

L'objectif de cette démarche expérimentale est de montrer qu'en partant d'un ensemble de données décrivant le domaine et d'un RB « dégradé » (c'est-à-dire qu'il capture la plupart des principales dépendances du domaine d'application, mais certaines sont manquantes ou erronées), il est possible de retrouver le RB qui représente le mieux le domaine d'application : en l'occurrence ici il s'agit du RB de référence que nous avons présenté dans la figure 3.8, page 74.

Dans la suite de cette section nous raisonnerons comme si RB_ref nous était inconnu, les seules « traces »que nous avons de ce RB sont des données générées. A partir du RB dégradé nous tenterons d'une part de retrouver quelles modifications ont été apportées à la structure et aux paramètres du RB, mais aussi de découvrir un

ensemble de règles d'association pertinentes sur le domaine.

3.7.2 Préparation du cas d'application

A partir de RB_ref, on produit un jeu de données composé de 10000 enregistrements. Ces données sont considérées comme étant les données du domaine *Visit Asia*. Comme nous cherchons à extraire des règles d'association on se focalise uniquement sur la *présence* des événements au sein des données.

Exemple. Si l'enregistrement est composé des attributs suivants : *Smoker* = *True*, *VisitAsia* = *False*, et *Dyspnea*=*Absent* alors on le codera uniquement par « *Smoker* ».

On modifie ensuite RB_ref –le RB qui a servi a générer les données– de telle sorte qu'on retrouve les différents cas de figures associés à l'étude d'un RB. Les modifications apportées sont les suivantes :

1. Le nœud *VisitAsia* n'est plus directement connecté au nœud *Tuberculosis*.
2. Du fait de la modification n°1 la tables de probabilités conditionnelles de *Tuberculosis* est modifiée de telle sorte qu'on a maintenant $p(Tuberculosis = Present) = 0,03$.
3. Les distributions de probabilités associées au nœud *Cancer* ont été modifiées afin que les valeurs de la variable *Smoking* n'influent pas sur *Cancer*. Plus précisément, RB_ref définissait

$$\begin{aligned} p(Cancer = Present|Smoking = Smoker) &= 0,1 \\ p(Cancer = Present|Smoking = NonSmoker) &= 0,01 \end{aligned}$$

   RB_0 définit maintenant

$$\begin{aligned} p(Cancer = Present|Smoking = Smoker) &= 0,1 \\ p(Cancer = Present|Smoking = NonSmoker) &= 0,1 \end{aligned}$$

4. La variable *Bronchitis* n'est plus directement reliée à *Dyspnea*.
5. Du fait de la modification n°4 les tables de probabilités de *Dyspnea* sont adaptées en conséquence.

Les changements apportés sont affichés en gras dans la figure 3.11, qu'il s'agisse de l'ajout d'un arc, ou de la modification d'une CPT. Ce RB correspond au réseau initial ou RB_0. Dans un contexte applicatif c'est ce réseau qui aurait été défini par un expert afin d'être utilisé en entrée de la $1^{\text{ère}}$ itération de notre processus.

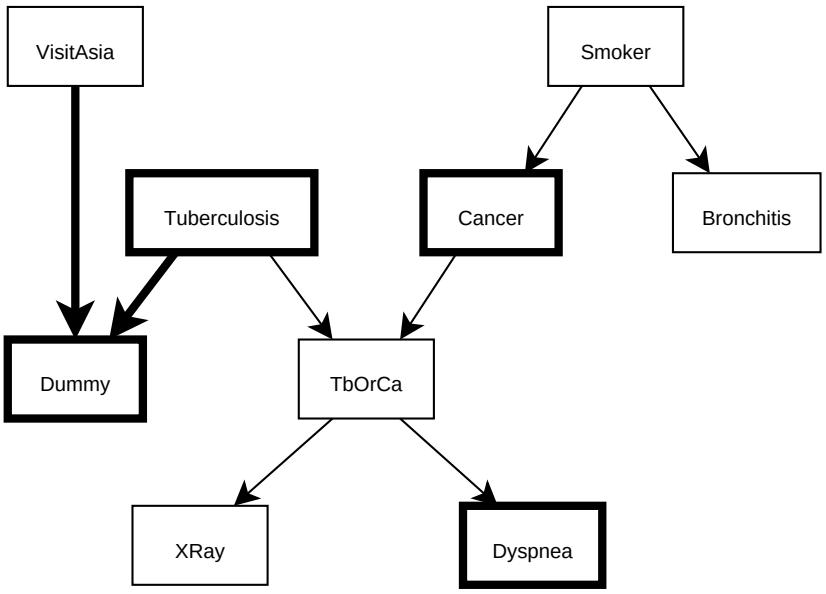

FIG. 3.11 – Réseau bayésien *Visit Asia* (RB_0) utilisé pour la 1ère itération du processus découverte de connaissances.

3.7.3 Déroulement de l'approche KARD

On se propose maintenant de suivre pas à pas les différentes étapes de la méthode présentée dans la figure 3.7.

Étape A

Le réseau *Visit Asia* RB_0 sert de base pour nos expérimentations. Nous avons décrit précédemment les modifications apportées par rapport à RB_ref.

Étape B

À partir du jeu de données générées, on extrait une collection concise de règles d'association ($minfreq = 100$, soit $0,01\%$ du nombre total d'enregistrements de la base de données et nombre maximal d'exceptions $\delta = 10$, i.e., ce qui nous garantit une confiance minimale de $0,90$). Il s'agit de la collection $\mathcal{L}\langle\mathcal{R}, \mathcal{I}_{\text{bd}}, \emptyset, \emptyset\rangle$. Les mesures objectives calculées sont la confiance et la moindre contradiction. Le temps d'exécution est négligeable, un total de 16 règles d'association sont extraites.

Étape C

À partir de $\mathcal{L}$ et de RB_0 on calcule ensuite $\mathcal{L}'\langle \mathcal{R}, \mathcal{I}_{\text{bd}}, \mathcal{I}_{rb\_0}, \mathcal{D}\text{-sep}\rangle$ qui comporte le résultat de la mesure d'intérêt vis à vis de RB_0, ainsi que l'ensemble des parties d-séparées des règles de $\mathcal{R}$.

Les résultats obtenus sont résumés dans le tableau 3.6. Avant de regarder de plus près à ces résultats, il faut se rappeler que les règles d'associations ne sont générées qu'à partir de la présence d'un événement particulier. Par exemple, la règle n°18 peut être lue de la façon suivante : « Si l'on observe que le patient est fumeur (*Smoker*), que l'on a effectué le diagnostic de la présence de dyspnée (*Dyspnea*) et d'une bronchite (*Bronchitis*), ainsi que l'activation du nœud spécifique (*TbOrCa*), alors des examens aux rayons-X révèlent souvent des résultats anormaux ».

Num	Règle d'association	conf	C-min	Int_{RB01}
1	Tuberculosis → TbOrCa XRay	0,98	0,17	0,01
2	VisitAsia → XRay	0,98	0,01	0,89
3	Bronchitis Cancer → TbOrCa XRay	0,97	0,47	0,01
4	Bronchitis TbOrCa → XRay	0,97	0,04	0,01
5	Cancer Dyspnea → TbOrCa	1,00	0,71	0,00
6	Cancer Smoking → TbOrCa	1,00	0,75	0,00
7	Cancer XRay → TbOrCa	1,00	0,80	0,00
8	Dyspnea Tuberculosis → TbOrCa XRay	0,97	0,16	0,01
9	Dyspnea VisitAsia → XRay	0,98	0,01	0,86
10	Bronchitis Cancer Dyspnea → TbOrCa XRay	0,97	0,41	0.01
11	Bronchitis Cancer Smoking → TbOrCa XRay	0,97	0,45	0.01
12	Bronchitis Dyspnea TbOrCa → XRay	0,97	0,03	0,01
13	Bronchitis Smoking TbOrCa → XRay	0,97	0,03	0,01
14	Cancer Dyspnea Smoking → TbOrCa	1,00	0,66	0,00
15	Cancer Dyspnea XRay → TbOrCa	1,00	0,69	0,00
16	Cancer Smoking XRay → TbOrCa	1,00	0,73	0,00
17	Bronchitis Cancer Dyspnea Smoking → TbOrCa XRay	0,97	0,40	0,07
18	Bronchitis Dyspnea Smoking TbOrCa → XRay	0,97	0,03	0,01
19	Cancer Dyspnea Smoking XRay → TbOrCa	1,00	0,64	0,00
20	Cancer → TbOrCa	1,00	0,84	0,00

TAB. 3.6 – Règles d'association extraites à partir de *Visit Asia* RB_01. Les items soulignés appartiennet à $\mathcal{D}$-sep(R).

Étape D

Dans la réalité l'expert dispose de l'outil d'aide à l'analyse que nous avons développé. Ici, nous allons devoir analyser les règles à partir du tableau 3.6. Dans ce tableau nous avons souligné les parties des règles qui appartiennent à $\mathcal{D}$-core, elles dénotent ce que l'on appelle les *dépendances principales* de la règle. Les parties de la règle qui ne sont pas soulignées contiennent, quant à elles, des informations qui n'ont pas été modélisées par le RB actuellement utilisé, ce sont les parties appartenant à $\mathcal{D}$-sep. En regardant ces résultats, deux règles d'association ont une valeur d'intérêt élevée : il s'agit des règles n°2 et n°9. La règle n°2 nous dit que : « Lorsqu'on observe qu'une personne a visité l'Asie alors on observe aussi des résultats de rayons-X anormaux ». Clairement cette règle apporte une information portée par les données, mais qui n'est pas modélisée en tant que *dépendance* par le RB. Ce fait est aussi corroboré par le fait que *XRay* est graphiquement séparé de *VisitAsia* : dans le réseau RB_0 l'information ne circule pas entre ces deux nœuds.

Ainsi il est possible de trouver des règles qui présentent une différence entre le modèle de connaissance disponible et les données. On peut néanmoins se demander si de telles découvertes d'associations sont réellement intéressantes, et si c'est le cas, quelles modifications peuvent être apportées au RB actuel pour refléter ces observations faites sur les données.

La phase d'annotation doit permettre de préparer une réponse à cette question. Dans notre cas, nous avons effectué nous même les annotations, au regard des valeurs d'intérêt et des décompositions des règles en parties d-séparées / dépendances principales. C'est en fait à ce moment qu'intervient le jugement de l'expert.

Prenons l'exemple des règles n°2 et 9 qui exhibent un intérêt élevé vis-à-vis de RB_0. Ces règles nous donnent une indication sur le sens de circulation de l'information entre *VisitAsia* et *XRay*. Mais si l'on se replace dans un contexte strictement médical, le fait de constater des examens de rayons-X anormaux n'est que la conséquence d'une maladie qui a été contractée. Après réflexion, il est donc plus judicieux de reformuler la règle découverte en une annotation pertinente reliant *VisitAsia* à *Tuberculosis*.

La collection d'annotations rédigée lors de cette étape est présentée dans le tableau 3.7.

Num	Annotations de l'expert	Règles impactées
1	(VisitAsia → Tuberculosis ; I)	{2, 9}
2	(Cancer → TbOrCa ; K)	{1, 5-7, 10, 11, 14-17, 19, 20}
3	(Tuberculosis → TbOrCa ; K)	{1, 8}

TAB. 3.7 – Annotations collectées sur les règles *Visit Asia*.

Les annotations n°2 et 3 témoignent d'un motif récurrent qui vient polluer la clarté des règles présentées. En effet la variable logique *TbOrCa* est activée chaque fois que *Tuberculosis=Present* ou *Cancer=Present* sont observés.

Étape E

Finalement, ces annotations sont transmises à l'expert chargé de la mise à jour du RB. L'examen des annotations permet de faciliter les éventuelles modifications de la structure et/ou des paramètres du RB.

On commence par s'intéresser à l'annotation n°1. En comparant ce motif avec la structure de RB_0, on remarque qu'aucun lien n'est modélisé entre *VisitAsia* et *Tuberculosis*. Ce motif étant marqué comme intéressant, l'expert en charge de la mise à jour du réseau va relier *VisitAsia* et *Tuberculosis* en assignant à la CPT de *Tubercolisis* une légère influence sur la présence de tuberculose lorsque le patient déclare avoir visité l'Asie.

Mais comment quantifier cette influence ? En effet il serait invraisemblable de modéliser le fait que tous les gens ayant effectué un voyage en Asie soient atteints de tuberculose ! Pour cela l'expert peut s'aider d'outils que nous avons mis à sa disposition, par exemple en étudiant, pour les règles concernées, le résultat des mesures objectives qui prennent en compte le nombre de contre-exemples (e.g., la moindre contradiction), ou encore en utilisant le module de test d'hypothèse : il est possible de tester à partir des données le support, et la confiance des règles : VisitAsia=NoVisit → Tuberculosis=Absent, VisitAsia=Visit → Tuberculosis=Absent, etc. Les résultats de cette analyse permettent d'établir la CPT associée au nœud *Tuberculosis*.

Nouvelle itération du processus

La première itération de notre processus est terminée. On peut alors initier une nouvelle itération sur les mêmes règles, mais cette fois à partir de RB_1, on obtient le résultat présenté dans le tableau 3.8.

En étudiant ce tableau on peut légitement se demander pourquoi l'intérêt des règles n°2 et 9 est toujours élevé. Il ne faut pas oublier que l'influence de *VisitAsia* sur *Tuberculosis* (et donc directement sur *XRay*) que nous avons modélisée, est – bien qu'intéressante médicalement – *faible*. Or ces deux règles font état d'une association *forte* en terme de confiance. Comme notre mesure est dépendante de la confiance, la valeur d'intérêt reste élevé, mais diminue par rapport à celle calculée sur RB_0.

Num	Règle d'association	Int_{RB_0}	Int_{RB_1}
1	Tuberculosis → TbOrCa XRay	0,01	0,01
2	VisitAsia → XRay	0,89	0,83
3	Bronchitis Cancer → TbOrCa XRay	0,01	0,01
4	Bronchitis TbOrCa → XRay	0,01	0,01
5	Cancer Dyspnea → TbOrCa	0,00	0,00
6	Cancer Smoking → TbOrCa	0,00	0,00
7	Cancer XRay → TbOrCa	0,00	0,00
8	Dyspnea Tuberculosis → TbOrCa XRay	0,01	0,01
9	Dyspnea VisitAsia → XRay	0,86	0,76
10	Bronchitis Cancer Dyspnea → TbOrCa XRay	0,01	0,01
11	Bronchitis Cancer Smoking → TbOrCa XRay	0.01	0,01
12	Bronchitis Dyspnea TbOrCa → XRay	0,01	0,01
13	Bronchitis Smoking TbOrCa → XRay	0,01	0,01
14	Cancer Dyspnea Smoking → TbOrCa	0,00	0,00
15	Cancer Dyspnea XRay → TbOrCa	0,00	0,00
16	Cancer Smoking XRay → TbOrCa	0,00	0,00
17	Bronchitis Cancer Dyspnea Smoking → TbOrCa XRay	0,07	0,01
18	Bronchitis Dyspnea Smoking TbOrCa → XRay	0,01	0,01
19	Cancer Dyspnea Smoking XRay → TbOrCa	0,00	0,00
20	Cancer → TbOrCa	0,00	0,00

TAB. 3.8 – Évolution de la mesure d'intérêt et des parties d-séparées calculées sur les règles d'association (RB_0 et RB_1

3.7.4 Critique des résultats obtenus

Par rapport aux objectifs que nous avions fixé Ces résultats mettent en avant plusieurs points suivants :

- La découverte de deux règles intéressantes a entraîné une modification du RB et a permit d'inverser
-
- A partir des règles étudiées il n'a pas paru possible de retrouver le lien de parenté entre *Bronchitis* et *Dyspnea* (modification n°1).
- De plus, comme nous nous intéressons uniquement à la présence des événements (e.g., le fait d'être fumeur) et non à leur absence (e.g., le fait d'être non fumeur), les règles présentées ne permettent pas de couvrir la modification n°3 réalisée sur la CPT de *Smoking*.

Le principal résultat a retenir ici est bien la découverte, à partir de l'élaboration d'un modèle – même incomplet – des dépendances du domaine, et son exploitation via la mesure d'intérêt subjective, de motifs réellement pertinents. On peut noter que dans les expériences que nous avons menées, aucune des mesures objectives utilisées ne permettait de retrouver facilement ces règles.

A titre de comparaison nous avons essayé une approche similaire mais cette fois en utilisant un algorithme de type Apriori, avec les mêmes contraintes et sur le même jeu de données. Au total 115 règles d'association sont générées. Parmi cette collection, trois règles mentionnent différentes variantes de la relation qui associe *VisitAsia* avec *XRay* et *Dyspnea*.

La principale différence entre notre approche et celle-ci plus naïve est que, dans le second cas, il va être beaucoup plus difficile de découvrir l'association intéressante qui implique l'attribut *VisitAsia*. Ceci est dû notamment au fait que beaucoup de règles – dont beaucoup sont redondantes – sont présentées à l'expert : celui-ci va devoir parcourir la totalité des règles avant de découvrir l'association pertinente.

Dans le prochain chapitre nous abordons un cas d'application plus complexe qui justifie les différents techniques et outils mis en place autour de notre processus.

Chapitre 4

Application à l'analyse des données d'interruptions opérationnelles

Au chapitre précédent nous avons décrit les contributions apportées sur le plan scientifique. Des expérimentations réalisées sur des données simulées nous ont permis une première validation de ces travaux de recherche. Ce chapitre est maintenant l'occasion de voir l'application de nos propositions sur une problématique réelle : l'aide à l'analyse des données d'interruptions opérationnelles dans l'aéronautique. Après un rappel et une présentation du cas d'application, nous déroulerons notre processus de fouille sur le jeu de données. L'analyse des résultats obtenus nous permettra de conclure sur l'efficacité de notre approche.

4.1 Description du cas d'application

Le développement d'un projet avion est actuellement basé sur les principes de l'*ingénierie concourante* afin de réduire autant que possible la durée du cycle de développement. Une des conséquences de cette approche est que les performances opérationnelles de l'avion doivent être estimées en amont du processus de conception, de telle façon que les exigences du client puissent piloter la conception du produit.

Une *interruption opérationnelle* (IO) arrive lorsqu'un problème technique (panne, disfonctionnement) empêche un avion de décoller lors d'une mission, au moins quinze minutes après l'heure de départ initialement fixée. Ces événements sont très importants pour les compagnies aériennes car les coûts engendrés sont loins d'être négligeables. Ainsi, très tôt dans le processus de conception de l'avion, les ingénieurs aéronautiques doivent réaliser une estimation réaliste de la fréquence des interruptions opérationnelles qui se vérifiera lorsque l'avion sera en opération. Ces prédictions – ainsi qu'un ensemble de contraintes spécifiques – initialisent, guident et valident les choix

de conception. Pour cela, les ingénieurs utilisent un outil qui implémente un modèle stochastique intégrant tous les paramètres qui sont connus pour avoir un impact sur la fréquence des interruptions opérationnelles. Cet outil est calibré et configuré à partir du retour d'expérience des avions en service dont les systèmes et les équipements ont des caractéristiques communes avec le projet en cours.

Actuellement les besoins de recherche tendent à se focaliser sur l'amélioration des modèles de calcul utilisés par l'outil de prédiction des performances opérationnelles. Dans ce contexte, la fouille de données en service est particulièrement intéressante puisqu'elle vise à découvrir des facteurs qui jusque là n'étaient pas connus. Ces facteurs pourraient alors être intégrés aux modèles pour obtenir des prédictions encore plus fidèles.

Les sections suivantes présentent l'application des différentes étapes du processus de fouille qui, selon la méthodologie proposée précédemment (Chapitre 3), aux données d'interruptions opérationnelles. L'objectif est de faciliter la découverte de règles d'association intéressantes et potentiellement exploitables – après reformulation – en tant que nouveaux contributeurs des taux d'interruptions opérationnelles.

Pour des raisons de confidentialité toutes les données présentées par la suite ont été maquillées de telle sorte que les références, les numéros d'ATA, les numéros de série, etc..., ne puissent pas être reconnus ou réutilisés à l'insu de la compagnie qui les détient.

Les travaux détaillés dans ce chapitre ont donné lieu à deux publications, une dans le domaine de l'extraction de connaissances [FDMB06a], l'autre plus axée sur les problématiques d'ingénierie de la connaissance [FDMB06b].

4.2 Mise en place du cadre de test

4.2.1 Description du jeu de données

Il existe différentes sources de données relatives aux interruptions opérationnelles. La base de données principale regroupe les détails de tous les problèmes techniques survenus en opération. Un extrait de cette base est présenté dans le tableau 4.1.

Le jeu de données est principalement composé d'attributs catégoriques comme le champ `EngineType` par exemple. On note aussi que le champ `Delay` est une valeur numérique qu'il va nous falloir discrétiser. Certains champs ne sont pas exploitables tels quels, d'autres nécessitent d'être enrichis : le champ `ATA` par exemple, désigne par un code à 6 chiffres un équipement de l'appareil. Ce chiffre fait partie d'une taxonomie utilisée par les ingénieurs aéronautiques. Par exemple l'ATA 212351 est un sous-ensemble du système 2123, qui est lui même un sous-système de la catégorie 21

ATA	Date	Opérateur	MSN	Moteur	Aéroport	Phase	Effet	Délai	Classe
0	29.12.1998	OP1	11	EngXXA	ST3	TX	DY	0.50	NM
0	30.12.1998	OP1	29	EngXXA	ST4	CS	DY	0.83	NA
212351	03.02.1998	OP2	11	EngXXA	ST4	CS	DY	0.68	
212600	07.10.1998	OP1	50	EngXXA	ST1	CS	DY	0.39	
212634	21.03.1998	OP2	142	EngXXA	ST4	TX	DY	0.85	
212634	23.03.1998	OP1	34	EngXXA	ST3	CS	DY	1.15	
212634	09.07.1998	OP1	87	EngXXA	ST3	CS	DY	0.25	
212634	04.09.1998	OP3	50	EngXXA	ST8	TO	DY	16.00	NM
212634	13.09.1998	OP4	42	EngXXA	ST2	CS	DY	2.37	
212651	07.09.1998	OP3	151	EngXXA	ST1	CS	DY	0.51	NS
212651	16.10.1998	OP5	170	EngXXA	ST3	CS	DY	0.42	

TAB. 4.1 – Extrait de la base de données d'interruptions opérationnelles.

(*système électromécanique*). L'expert souhaite qu'une analyse du numéro ATA puisse s'effectuer à ses différents niveaux de décomposition (i.e. à 2, 4 et 6 chiffres).

Enfin, une des spécificités de ce jeu de données est la présence d'un champ de texte libre qui décrit l'incident et donne – éventuellement – des détails supplémentaires sur l'interruption opérationnelle : quelles pannes ont été détectées ? Quelles actions ont permit la remise opérationnelle de l'appareil. Les informations contenues dans ce texte libre ne sont malheureusement pas toutes reportées dans les autres champs de la base. De plus ces textes étant rédigés par des opérateurs de maintenance différents, sans formalisme imposé, leur structure et leur sémantique n'est donc pas homogène. Un exemple de texte libre est présenté dans la Figure 4.1.

```
Troubles with entertainment system after installation of new software.
Crew tried reset nil fix.
EPESC (enhanced pax entertainment system controller) replaced.
Old software loaded.
System repaired by MAS representant.
```

FIG. 4.1 – Exemple de texte détaillant une interruption opérationnelle

Enfin il faut aussi savoir que les données utilisées sont relativement bruitées et que quelques données sont manquantes (champs non renseignés). La quantité de bruit est cependant difficile à estimer. Le nombre d'enregistrement incomplets étant faible nous avons simplement décidé de retirer ces données du jeu de départ.

L'expert souhaitant concentrer ses recherches sur une famille de produit homogène nous avons limité la base de données à 11819 enregistrements correspondants au programme avion sélectionné par l'expert.

4.2.2 Pré-traitements

Les pré-traitements à effectuer sur ces données sont de trois ordres :
- Enrichissement d'un attribut, c'est-à-dire créer un ou plusieurs attributs qui vont enrichir ou remplacer l'attribut en question, par exemple en croisant les données avec d'autres tables.
- Modification, simplification des valeurs prises, selon des règles précisées par l'expert.
- Discrétisation, à partir des attributs catégoriques ou numériques.

Pour effectuer ces pré-traitements nous avons utilisé uniquement les requêtes SQL. Cela nous a permis d'automatiser les traitement pour, à la fois convertir certains champs de la base initiale, et pour croiser les données avec d'autres tables. Cette approche à l'avantage d'être relativement flexible dans le cas où on voudrait modifier les données utilisées.

Exemple. Les commandes SQL ci-dessous nous permettent d'appliquer des règles définies par l'expert pour la préparation des différents champs de la base. A titre d'illustration on ne présente qu'un extrait de la requête globale (Figure 4.2.2).

```
 [...]
UPDATE operational_interruption
SET effect='DY'
WHERE effect IS NULL AND code_effect LIKE '%DY%';

UPDATE operational_interruption
SET effect='CN'
WHERE effect='DY' AND delay>=6;

UPDATE operational_interruption
SET  delay_interval = '0.0_0.5'
WHERE  delay<0.5;
[...]
```

FIG. 4.2 – Extrait de la requête SQL pour le pré-traitement des données.

Ici, ces commandes ont pour effet d'affecter une valeur particulière à l'attribut *effect* selon la composition de la variable *code_ effect* et de discrétiser la variable *delay* en différents intervalles spécifiés par l'expert.

Le champ de texte libre a quant à lui bénéficié d'un traitement particulier (Section 4.2.3). L'objectif étant d'en retirer le maximum d'information, sous la forme de nouveaux attributs que l'expert souhaite voir apparaître dans les règles d'association.

À l'issue de la phase de pré-traitement, on dispose de 21 attributs discrétisés au total en 2004 valeurs différentes et de 11819 enregistrements. La matrice d'entrée est donc de taille : 2004 colonnes par 11819 lignes. La taille moyenne d'un enregistrement (c'est-à-dire le nombre moyen d'événements observés) est de $14,6$. Cet ensemble de données est jugé suffisant par l'expert du domaine pour travailler sur la recherche de nouveaux facteurs ayant un impact sur la fréquence des interruptions opérationnelles.

Tous les calculs ont été effectués à partir d'un ordinateur standard (processeur 2 GHz, 1 Go de mémoire).

4.2.3 Exploitation du texte libre

Une analyse manuelle de plusieurs champs de textes libres ainsi qu'une discussion avec l'expert a mis en avant l'importance de ce champ dans l'étude des données d'IO. Il est effectivement porteur d'informations intéressantes pour l'expert. Ceci nous a poussé à mettre en place une analyse assez fine de son contenu. L'objectif étant de pouvoir extraire certaines caractéristiques de ce texte, lorsqu'elles sont présentes.

Le sujet de la thèse n'étant pas la fouille de texte, on restera toutefois assez brefs quant aux techniques employées. On ne prétend pas non plus que la méthode utilisée est la plus adaptée à notre cas d'application.

La démarche est simple, on va chercher à extraire de chaque texte différentes informations relatives au « contexte » de l'IO au « problème »effectivement constaté par les équipes de maintenance, ainsi qu'aux « actions » qui ont été menées pour tenter de remettre l'avion en opération. À chacun de ces champs correspond un ensemble de mots clés dans le texte libre et un ensemble d'attributs que l'on souhaite utiliser pour la fouille de données.

Grâce à un ensemble de règles fournies par notre expert, nous avons établi une chaîne de traitement capable d'effectuer ce découpage et d'extraire les mots clés pertinents. De plus, nous avons décidé d'un commun accord avec l'expert d'accorder une importance particulière à la « dernière action » réalisée car elle est bien souvent synonyme d'action correctrice par rapport au problème initialement détecté.

Exemple. Si l'on repart du texte présenté dans la Figure 4.1 on va détecter qu'un problème est intervenu sur un système particulier (*troubles with entertainment system*). Ici il n'y a pas d'informations supplémentaires par rapport au numéro d'ATA qui accompagne le texte. Ensuite nous détectons deux interventions de l'équipage : une remise à zéro de l'équipement qui ne permet pas de résoudre le problème (*reset nil fix*), puis un remplacement de l'équipement incriminé (*EPESC replaced*). Cela nous permet de remplir le champ « actions » en mettant les mots-clés *reset* et *replace* à *vrai*. La dernière action est détectée comme étant hors contexte, puisqu'il s'agit de la réparation de l'équipement en dehors du cycle opérationnel de l'avion. On assigne

donc au mot clé *last_ action* la valeur *replace*.

L'intérêt de cette démarche est évident pour l'expert, car cela permet de faire intervenir, pour chaque enregistrement, plus de précisions sur l'interruption opérationnelle. En effet, pour un problème donné il est évident que la *pose-dépose* d'un équipement a plus de chances d'entrainer une interruption opérationnelle de longue durée que la simple remise à zéro de l'équipement incriminé.

4.3 Expérimentations réalisées

Notre démarche expérimentale suit le processus de fouille que l'on a présenté précédemment. Pour rappel il se décompose en différentes étapes :

1. Définition de la structure initiale du réseau bayésien.
2. Génération d'un ensemble concis de règles d'association.
3. Calcul de l'intérêt et des parties d-séparées des règles vis-à-vis du réseau bayésien.
4. Étude des règles et annotation par l'expert
5. Mise à jour de la structure et des paramètres du réseau bayésien à partir des annotations.

Nous allons étudier le déroulement de ces différentes étapes sur les données d'IO.

4.3.1 Définition du réseau bayésien initial

La première étape consiste donc à modéliser, en utilisant le formalisme propre aux Réseaux Bayésiens, les principales dépendances du domaine.

Afin de préparer cette modélisation initiale, nous avons commencé par sensibiliser l'expert des données IO aux Réseaux Bayésiens. Ainsi nous avons fait la démonstration de la circulation de l'information dans le graphe, présenté le mécanisme d'inférence à partir d'exemples concrets et expliqué la signification et la définition des tables de probabilités conditionnelles associées aux nœuds du graphe.

Enfin nous avons demandé à l'expert de modéliser les dépendances du domaine qui lui paraissaient importantes. Plus précisément nous nous sommes concentrés sur la définition d'une première structure du RB par l'expert. Cependant, il est relativement délicat d'interpréter la structure « seule » du RB, de ce fait nous avons donné pour consignes à l'expert d'établir un lien de causalité entre deux variables A et B lorsqu'il pensait que « la connaissance de la valeur prise par A apportait une connaissance supplémentaire sur les valeurs que pouvaient prendre B ».

Le résultat obtenu est présenté dans la Figure 4.3. On part du principe que ce première modèle capture certaines dépendances du domaine d'application, d'autres restent à préciser ou à découvrir grâce à notre approche.

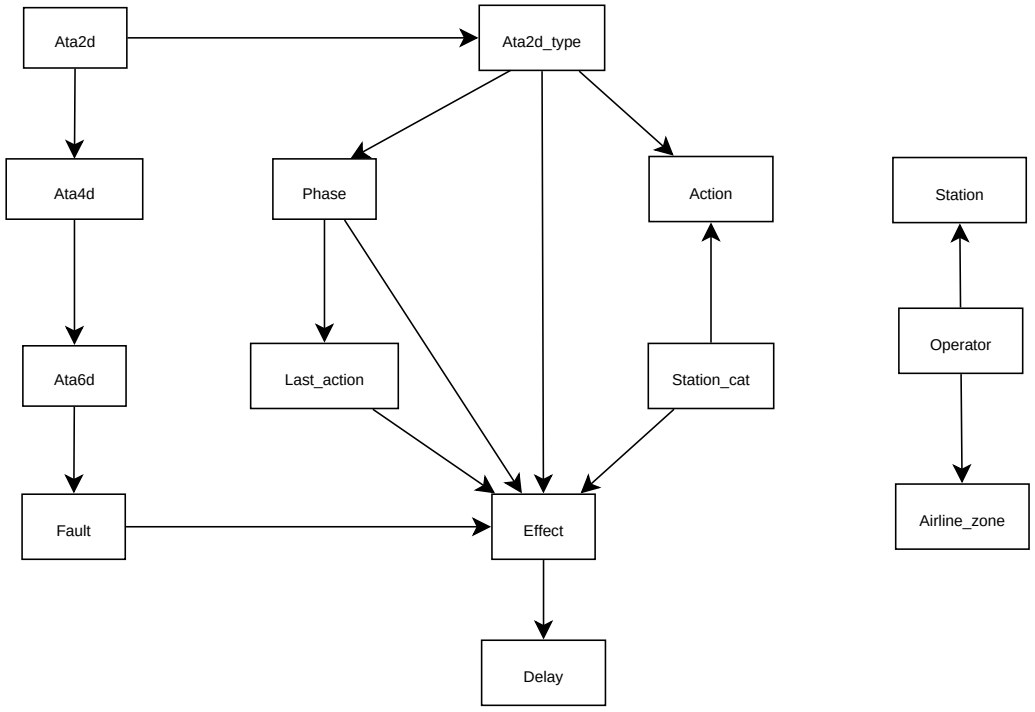

FIG. 4.3 – Réseau bayésien initial (RB01) sur les données IO.

A partir de cette structure et des données d'IO nous effectuant un apprentissage automatique des tables de probabilités conditionnelles grâce à l'algorithme d'apprentissage implanté dans l'API Weka [WF05].

4.3.2 Génération d'un ensemble concis de règles d'association

Pour calculer une collection de règles d'association non redondantes nous utilisons AC-LIKE une implémentation de l'algorithme AC-MINER [BBR00]. Les paramètres utilisés sont $Freq = 100$ et $\delta = 20$). Sur notre jeu de données, le temps d'extraction est inférieur à la minute. Au total, 5633 itemsets δ-libres fréquents ainsi que leur δ-fermeture sont extraits. A partir de cet ensemble nous générons 4954 règles δ-fortes. Ces règles ne vont pas être chargées en totalité dans notre application. Afin de faciliter l'étape d'analyse par l'expert, celui a la possibilité d'établir des filtres syntaxiques, de seuiller selon les différentes mesures d'intérêts calculées, etc.

4.3.3 Exploitation du réseau bayésien sur les règles extraites

Num	Règle d'association	Confiance	Int_{RB01}
1	last_action=swap → swap	1,00	0,96
2	delay<1h last_action=swap → DY swap	1,00	0,96
3	effect=DY last_action=swap → swap	1,00	0,96
4	last_action=swap CS → DY swap	0,94	0,92
5	zone=E last_action=swap CS → DY swap	0,95	0,92
6	Electro-Mechanical last_action=swap CS → DY swap	0,94	0,91
7	delay<1h last_action=swap CS → DY swap	1,00	0,95
8	last_action=mel OP1 MB → zone=E DY mel ST1	0,99	0,94
9	last_action=swap CS MB → DY swap	0,95	0,92
10	Electro-Mechanical last_action=nff CS MB → DY nff	0,99	0,91

TAB. 4.2 – Règles ayant la plus forte valeur d'intérêt vis-à-vis de RB01.

On calcule la mesure d'intérêt des règles d'association par rapport au RB utilisé pour ce premier cycle du processus. Pour cela, un algorithme Kruskal basé sur la réduction en polyarbres [Chr75], implémenté dans la bibliothèque *Bayesian Network Tools in Java*[1], a été utilisé pour réaliser des calculs d'inférence approchés. Ce processus a duré approximativement 20 heures pour traiter l'ensemble des 4954 règles, soit un temps moyen de 14,55 secondes par règle. Le temps de calcul peut paraître long mais il faut savoir qu'au moment où les calculs ont été réalisés, aucune optimisation n'était utilisée.

En particulier il n'y a pas de système de « cache » qui permettrait d'économiser de nombreux calculs d'inférence. Prenons l'exemple de deux règles (1) A=a B=b → X=x et (2) A=a B=b C=c → X=x Y=y. Notre calcul d'intérêt implique que nous calculions, pour ces deux règles, la distribution de probabilité $p(X)$, sachant la partie droite de la règle. Maintenant admettons que pour la règle (2) la variable $C = c$ n'intervienne pas dans le calcul de $p(X)$, alors on va effectuer deux fois le même calcul d'inférence.

Le tableau 4.2 montre les 10 premières règles, classées selon la valeur de leur mesure d'intérêt par rapport au RB initial (ou RB01). Ces résultats font intervenir des codes et des sigles propres au domaine des IO ce qui les rend difficile à interpréter. Nous allons expliciter quelques règles pour faciliter la compréhension générale des résultats présentés dans ce chapitre.

Par exemple, a règle « **zone=E last_action=swap CS → DY swap** », peut se lire de la façon suivante : « Si la compagnie qui opère l'avion est en zone européenne (*zone=E*), que le problème survient pendant la phase de vérification de l'appareil

[1] `http://sourceforge.net/projects/bnj`

(*check list* ou *CS*) et que la dernière action effectuée par l'équipe de maintenance est un échange standard de l'équipement incriminé (*last_ action=swap*), alors on observe très souvent un délai inférieur à six heures (*DY*) et le fait qu'un échange standard a été effectué (*swap*) ».

Prenons un autre exemple, ainsi **CS ST3 → zone=ME DY OP3 other**, nous dit que : « Si le problème à lieu au moment de la phase de vérification au sol (*CS*) et que l'on se trouve à l'aéroport désigné par le sigle *ST3*, alors on observe souvent l'ensemble des faits suivant : on se trouve au moyen-orient (*zone=ME*), l'interruption est un retard inférieur à six heures (*DY*), la compagnie qui opère l'avion est désignée par *OP3*, et l'aéroport où a lieu l'interruption n'est ni la base principale de la compagnie (*MB*), ni celle d'une autre compagnie (*other*) ».

Enfin les codes à 2, 4 ou 6 chiffres représentent l'ATA de l'équipement incriminé dans l'IO à différents niveaux de décomposition. Pour une raison de confidentialité tous les numéros ont été maquillés.

4.3.4 Étude des règles d'association et annotation

L'étape suivante de notre processus consiste, pour l'expert du domaine, à étudier les règles générées, puis proposer des annotations en utilisant la syntaxe adéquate.

Dans un premier temps les règles sont classées selon la mesure d'intérêt calculée par rapport au RB. On décide, arbitrairement, de fixer la contrainte d'intérêt minimal à $0,75$: 408 règles satisfont cette contrainte.

L'expert constate immédiatement un « problème » avec ces premiers résultats : en effet le lien last_action → action n'a pas été modélisé dans le RB, alors qu'il s'agit d'une relation logique entre deux variables. En conséquence, toutes les règles faisant intervenir ce lien sont jugées intéressantes. C'est parfaitement en accord avec le fonctionnement attendu du RB. Comme il s'agit d'une modification importante l'expert décide d'annoter directement cette relation comme étant une relation connue (K) avec pour consigne de modifier la structure du RB de telle façon que ces annotations soient prises en compte.

Num	Annotations de l'expert	Règles impactées
1	(last_action=swap → swap ; K :'certain')	{1–7, 9}
2	(last_action=mel → mel ; K :'certain')	{8}
3	(last_action=nff → nff ; K :'certain')	{10}

TAB. 4.3 – Exemple d'annotations collectées à la première itération du processus.

Un examen montre que 1335 règles ont un intérêt proche de zéro ($\text{Int}_{RB01} < 0,1$) et peuvent être éliminées car elles représentent toutes, sans exceptions, une informa-

tion évidente ou déjà connue de l'expert. A la première itération du processus cela représente 27% de la collection de règles générée.

On peut aussi noter, comme on l'avait remarqué pour l'exemple *Visit Asia* présenté au Chapitre 3, qu'il n'y a aucune corrélation entre notre mesure et les autres mesures d'intérêt (confiance, j-mesure, moindre contradiction).

4.3.5 Mise à jour du réseau bayésien

Les annotations précédemment collectées (résumées dans le tableau 4.3) ont été passées à un autre expert dont la tâche était d'apporter des modifications au RB. Ici, nous avons décidé de faire une intégration directe de toutes les annotations (K). Les tables de probabilités ont également été modifiées en conséquence. Pour des raisons évidentes – de place, et de confidentialité – elles ne peuvent pas être présentées ici. Le lecteur peut néanmoins se reporter à la Figure 4.4 pour évaluer les modifications apportées (en gras dans le graphe.)

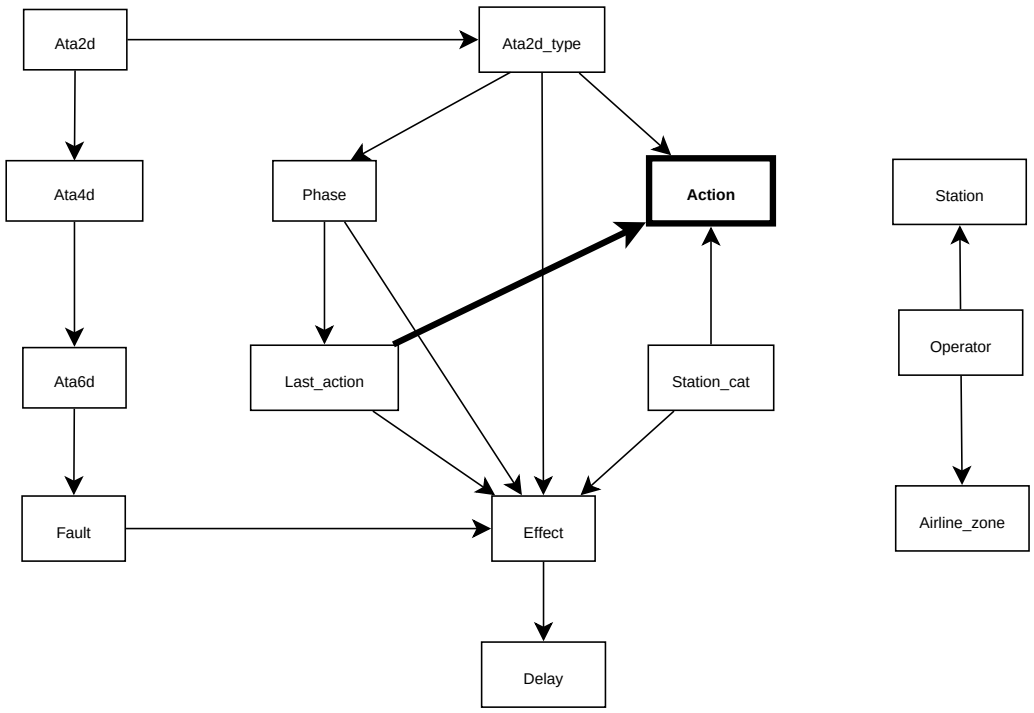

FIG. 4.4 – Réseau bayésien à l'issue de la première mise à jour (RB02).

4.3.6 Nouvelles itérations du processus

Itération n°2

Après la mise à jour du RB, une nouvelle itération du processus est initiée. On retourne donc à l'étape 3, c'est-à-dire au calcul des mesures d'intérêt par rapport au nouveau RB. Le tableau 4.4 montre l'évolution de la mesure d'intérêt avant et après la mise à jour du RB, sur le même ensemble de règles. Il apparaît clairement que les modifications effectuées permettent de filtrer toutes les règles qui avaient été jugées intéressantes (Int $> 0,75$) lors de la première itération.

Num	Règle d'association	Int_{RB01}	Int_{RB02}
1	last_action=swap → swap	0,96	0,03
2	delay<1h last_action=swap → DY swap	0,96	0,02
3	effect=DY last_action=swap → swap	0,96	0,02
4	last_action=swap CS → DY swap	0,92	0,31
5	zone=E last_action=swap CS → DY swap	0,92	0,32
6	Electro-Mechanical last_action=swap CS → DY swap	0,91	0,25
7	delay<1h last_action=swap CS → DY swap	0,95	0,02
8	last_action=mel OP1 MB → zone=E DY mel ST1	0,94	0,50
9	last_action=swap CS MB → DY swap	0,92	0,26
10	Electro-Mechanical last_action=nff CS MB → DY nff	0,91	0,21

TAB. 4.4 – Évolution de la mesure d'intérêt avant et après modification (RB01 et RB02).

Le tableau 4.5 présente quant à lui les dix meilleures règles, selon la nouvelle valeur d'intérêt. Des 408 jugées intéressantes à la première itération il ne reste plus que 39 règles telles que $\text{Int}_{RB02} > 0,75$. En fait ce seuil nous a servi dans le but de composer un jeu de règles étalon que l'on va suivre tout au long du processus de fouille ; l'expert, lui, peut s'intéresser à un ensemble un peu plus vaste de règles (en pratique les règles dont l'intérêt est supérieur à $0,25$).

Cette nouvelle phase d'étude des règles permet à l'expert de commencer à découvrir des relations d'association. Cependant avant d'annoter un motif comme étant réellement intéressant (I) l'expert doit être en mesure de vérifier que le motif en question n'est pas le fait d'une simple coïncidence sur les données. Pour ce faire il a à sa disposition des outils de tests qui vont lui permettre d'évaluer plus finement le motif en question sur l'ensemble données.

Exemple. Prenons la règle **A=a B=b → C=c**. L'expert pense que le motif **A=a → C=c** est potentiellement intéressant. Afin de pouvoir confirmer la validité de ce motif il doit évaluer le comportement de la règle lorsque les variables B=b et C=c ne

Num	Règle d'association	Int_{RB01}
1	zone=ME delay<1h CS MB → DY OP2 ST2	0,87
2	801120 → 80 8011 DY	0,87
3	4900 → Engine 49 490000 DY CS	0,85
4	2851 nff → Electro-Mechanical 28 285134 DY CS	0,83
5	CS ST3 → zone=ME DY OP3 other	0,82
6	Avionic smoke → 26 DY	0,80
7	4900 delay<1h → Engine 49 490000 DY CS	0,79
8	delay<1h OP4 ST4 → zone=NA DY other	0,79
9	zone=ME last_action=remove MB → DY remove OP2 ST2	0,77
10	2851 delay<1h → Electro-Mechanical 28 285134 DY CS	0,76

TAB. 4.5 – Règles d'association ayant la plus forte valeur d'intérêt vis-à-vis de RB02.

sont pas présentes. Pour ce faire il étudie l'évolution des mesures de confiance et de moindre contradiction dans les différents cas de figure (i.e. « A=a B=b → C=c », « A=a B=b ⇒ C=¬c », « A=a B=¬b ⇒ C=c », et « A=a B=¬b ⇒ C=¬c »).

Une partie des annotations réalisés lors de cette itération est donnée dans le tableau 4.6.

Certains motifs (sous-partie d'une règle d'association) sont jugés non valides par l'expert. C'est le cas par exemple de la relation zone=ME → OP2 ST2 (ou de manière plus générale : zone=X → OP ST), qui nous dit que lorsque l'on connaît la zone géographique dans laquelle opère une compagnie, alors on peut déduire le nom de cette compagnie et l'aéroport où a eu lieu l'interruption. Ce type de relation n'a pas caractère de connaissance valide selon l'expert, il s'agit d'une spécificité du jeu de données pour le triplet (zone=ME, OP2 ST2). Les annotations correspondantes sont alors rédigées dans la catégorie (NV).

D'autres motifs sont annotés comme étant connus (K) car ils représentent une relation évidente. En l'occurrence ces relations sont souvent présentes à l'intérieur d'une règle et déjà modélisées par le RB. L'annotation systématique de ces motifs va se répercuter sur l'affichage des règles. Le but est de permettre à l'expert une visualisation immédiate des parties de la règle qui représentent une information déjà connue par rapport à celles qui n'ont pas été annotées. C'est le cas par exemple pour toutes les relations du type ata4d → ata2d, ata6d → ata2d ata4d, etc.

Un autre exemple d'annotation de type (K) est l'annotation operator station → station_cat qui représente une information qui n'est pas modélisée dans le RB, mais qui est connue de l'expert.

L'expert remarque que la règle n°2 (tableau 4.5) a une valeur d'intérêt élevée alors qu'elle représente une information qui, a priori, est définie dans le RB. Cela est dû au

fait que nous avons réalisé un apprentissage automatique des tables de probabilités jointes. L'algorithme utilisé ayant naturellement tendance à lisser la distribution des probabilités, certains cas peuvent paraître intéressants alors que l'expert à explicitement modélisé la dépendance qu'ils représentent. Pour y remédier, une annotation de type (NP) est rédigée concernant cette règle. A la différence des annotations de type (NV), le motif représente bien une dépendance valide mais celle-ci ne fait pas sens par rapport aux objectifs de recherche de l'expert. On ne souhaite pas non plus réévaluer les probabilités associées à la variable *Ata6d=801120*, car la règle n°1 ne fait que mettre une spécificité du jeu de données utilisé. Cette particularité n'a pas lieu d'être explicitement prise en compte dans le RB, on choisi donc de signaler ce motif comme étant « non pertinent ». On verra à l'itération n°3 de notre processus que les annotations de type (NP) ont un impact visuel lors de l'affichage des règles d'association.

Enfin, on constate que les motifs du type **ata2d → ata6d** sont jugés intéressants. En effet ce type de relation va à l'encontre de ce qui est modélisé dans le RB (i.e. si par définition il est certain d'avoir la relation ata6d → ata2d l'inverse n'est a priori pas vrai). Il s'agit en fait d'une particularité intéressante du jeu de données. L'exploitation de ce type de découvertes est alors laissée à la discrétion de l'expert du domaine.

Num	Annotations de l'expert	Règles impactées
1	(zone → OP ; NV)	{1, 9}
2	(zone → ST ; NV)	{1, 9}
3	(ata6d=801120 → Engine ; NP)	{2}
4	(ata6d=801120 → ata2d=80 ; NP)	{2}
5	(ata6d=801120 → ata4d=8011 ; NP)	{2}
6	(operator station → station_cat= ; K :'certain')	{8}
7	(ata2d → ata_type ; K :'certain')	{}
8	(ata4d → ata6d ; I)	{3, 4, 7, 10}
9	(ata4d action → ata6d ; I)	{4}
10	(ata_type fault → ata2d ; I)	{6}
	...	...

TAB. 4.6 – Exemples d'annotations collectées à la deuxième itération du processus.

Les annotations sont ensuite utilisées pour effectuer la mise à jour du RB. Le résultat est présenté dans la Figure 4.5, les modifications apparaissent en gras.

Itération n°3

L'expert peut décider de visualiser les nouveaux calculs d'intérêt qui découlent de l'utilisation de RB03. A titre d'illustration on présente là aussi les dix premières règles

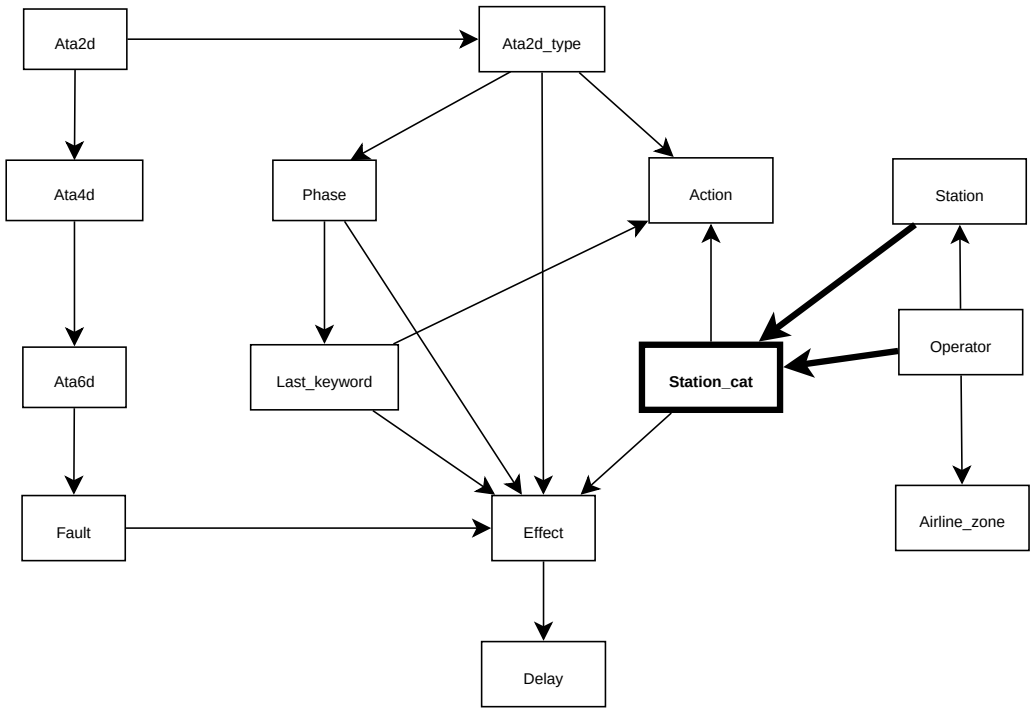

FIG. 4.5 – Réseau bayésien à l'issue de la deuxième mise à jour (RB03).

obtenues. Cette fois on applique visuellement l'impact des annotations (K), (NP) et (I) sur les règles (tableau 4.6). Nous avons choisi ici d'utiliser un code typographique spécifique à chaque type d'annotation. Ainsi les parties de la règle qui correspondent à une annotation seront mises en italique, barrées ou soulignées, si elles sont respectivement, déjà connues, non pertinentes ou intéressantes. L'application d'un code visuel aux règles, en fonction des annotations, permet à l'expert de se représenter plus rapidement quelles sont les différentes informations apportées par les règles. Le résultat est présenté dans le tableau 4.7.

Cette troisième itération nous confirme les règles qui avaient été jugées intéressantes à l'étape précédente. De plus, comme pour le passage de l'itération n°1 à la n°2, cette nouvelle mise à jour du RB élimine les règles déjà connues.

4.4 Critique des résultats obtenus

Nous avons appliqué notre processus de découverte de connaissances aux données d'IO dans le domaine aéronautique. Le réseau bayésien finalement obtenu est présenté dans la Figure 4.5. En le comparant avec le réseau construit initialement (Figure 4.3), on s'aperçoit que notre processus a permis l'ajout de plusieurs arcs qui n'avaient pas

Id	Règle d'association	Int_{RB03}
1	~~801120~~ ⇒ ~~Engine~~ ~~80~~ ~~8011~~ DY	0,87
2	4900 ⇒ *Engine 49* 490000 DY CS	0,85
3	2851 nff ⇒ *Electro-Mechanical 28* 285134 DY CS	0,83
4	7200 DY ⇒ *Engine 72* 720000	0,80
5	Avionic smoke ⇒ 26 DY	0,80
6	4900 delay<1h ⇒ *Engine 49* 490000 DY CS	0,79
7	2851 delay<1h ⇒ *Electro-Mechanical 28* 285134 DY CS	0,76
8	2793 *last_ action=remove* ⇒ *Avionic 27* 279334 DY *remove*	0,74
9	zone=AP last_action=nff CS ⇒ delay<1h DY nff	0,52
9	Electro-Mechanical delay>6h CS remove ⇒ CN last_action=remove	0,52
10	mel ST1 ⇒ zone=E DY last_action=mel OP1 MB	0,50

TAB. 4.7 – Règles d'associations évaluées par rapport à RB03 et aux annotations.

été « trouvés ». C'est le cas par exemple de la dépendance logique entre les nœuds *Last_ action* et *action* mais aussi pour les liens entre *operator, station* et *station_ cat.* Cela montre qu'un « oubli » lors de la modélisation initiale est rapidement détecté par notre système, et il peut être corrigé très facilement.

De plus, les précisions successivement apportées au RB ont non seulement permis un filtrage de plus en plus efficace dès règles connues, évidentes ou non intéressantes, mais cela a aussi clairement facilité la découverte de règles pertinentes. Ainsi, à la troisième itération de notre processus les règles ayant un fort intérêt par rapport au RB sont effectivement porteuses d'informations jugées intéressantes par l'expert.

Le deuxième avantage de notre approche est la possibilité d'éliminer les règles qui ne présentent pas d'intérêt pour l'expert. Un aspect rassurant vis-à-vis du filtrage des règles à très faible valeur d'intérêt (seuil arbitrairement fixé à $\text{Int}_{RB} < 0,1$) provient du fait que, sur nos expériences, aucun faux-négatif n'a été détecté. C'est-à-dire qu'il n'y avait pas de règles que l'on aurait pu juger intéressantes parmi les règles filtrées.

Au final, la découverte de ces dépendances et les modifications qui en ont découlées confortent le principe proposé par notre approche qui est de modéliser, par interaction avec l'expert, les dépendances du domaine et les exploiter pour améliorer le filtrage des règles d'associations extraites et ainsi faciliter leur étude.

Ces résultats soulèvent néanmoins quelques nouvelles interrogations. La première vient du fait que certaines règles qui possèdent un motif non valide peuvent avoir une valeur d'intérêt élevée. C'est notamment le cas de la règle d'association n°1 présentée dans le tableau 4.7. Il s'agit de cas de « faux-positifs ». La solution actuellement apportée est uniquement visuelle : la sous-partie de la règle qui contient une information non valide est présentée différemment à l'expert. Il faut donc réfléchir au filtrage de

ces motifs autrement que par la mesure d'intérêt.

Un deuxième problème a prendre en compte dans des travaux futurs est l'étude de l'impact potentiel que peut avoir la modification de la structure ou des paramètres du RB, sur l'ensemble règles. Il s'agit en quelque sorte de pouvoir contrôler et mesurer les effets de bord provoqués par la mise à jour du RB. Il peut par exemple être intéressant de présenter à l'expert uniquement les règles qui ont été impactées par la modification du modèle.

On a aussi vu que les calculs de mesures d'intérêt effectués dans ce chapitre pouvaient être longs lorsqu'on travaille sur des jeux de données réels, relativement complexes. En fait, il faut savoir que le temps de calcul est linéaire en fonction de la taille du RB (nombre de nœuds, nombre d'arcs), du cardinal moyen des itemsets de la partie gauche des règles et bien sûr du nombre total de règles à traiter. Avec la mise en place d'un système de cache approprié il est envisageable de réduire le temps de calcul total de plusieurs ordres de grandeurs. N'ayant pas de contraintes de temps de calculs spécifiques pour notre application nous n'avons pas cherché à creuser cette voie.

Chapitre 5

Conclusion et perspectives

Principales contributions

Dans ce mémoire nous avons proposé une approche pour faciliter la tâche d'analyse des résultats d'extraction de règles d'association. Nous avons d'abord présenté, dans le chapitre 1, le cadre général de la découverte de connaissances, ainsi que le contexte industriel qui a motivé les travaux de thèse. Dans un processus opérationnel, le *modèle* et les *données* évoluent tous les deux dans le temps. On constate généralement un décalage entre ces deux entités. Parfois les différences sont connues des experts, parfois elles restent à identifier. Pour ce deuxième cas de figure, la mise en place de techniques de découverte de la connaissances est intéressante. En particulier nous nous sommes intéressés à l'utilisation d'un réseau bayésien pour faciliter la découverte de règles d'association pertinentes sur de grands volumes de données.

Le deuxième chapitre a été l'occasion de réaliser un état de l'art sur les techniques actuelles pour l'extraction de règles d'association. Nous avons mis en avant les problèmes ouverts sur le domaine, à savoir les limites des approches dites « objectives » pour la découverte de règles réellement intéressantes. Nous avons aussi vu que, malgré les récentes propositions sur cette problématique, il n'y avait pas de prise de recul visant à établir une réflexion sur une utilisation plus *systématique* des connaissances du domaine, ainsi que sur la définition du rôle de l'expert au sein du processus de découverte.

Dans le troisième chapitre nous avons choisi de positionner nos contributions sur trois axes. Tout d'abord sur le domaine des collection de règles d'association non redondantes. Nous avons étudié les propriétés des ensembles de règles générées à partir des itemsets (δ-)libres et de la δ-fermeture. Cette étude nous a permis d'établir une base pour la génération de règles approximatives (i.e., dont ne peut déterminer précisément la confiance). Le deuxième axe de contribution concerne notre apport au domaine de l'ingénierie de la connaissance, en particulier sur l'étude des inter-

actions avec le modèle de connaissance utilisé pour faciliter la découverte de règles pertinentes (construction, exploitation, évolution). Nous avons aussi montré l'importance du rôle de l'expert dans ce processus, notamment par le biais des annotations sur les règles d'association. En effet, nous avons appliqué notre approche sur un cas d'application concret de l'industrie aéronautique, l'aide à l'analyse de données d'interruptions opérationnelles. Ce chapitre apporte une confrontation de nos contributions à des problèmes concrets de l'industrie.

Génération d'un ensemble non redondants de règles d'association à la confiance contrôlée

Notre première contribution a consisté à étudier les propriétés des collections de règles non-redondantes actuelles et de voir s'il était possible d'aller « plus loin »en sacrifiant une petite part de précision sur la confiance des règles. Pour cela nous sommes partis des travaux de [BBR00, Pas00] et nous avons vu qu'il était possible de définir deux nouvelles bases pour la génération de règles d'association dont la confiance est bornée par un paramètre δ. Les règles ainsi générées ont été utilisées sur notre cas d'application aux données d'interruptions opérationnelles.

Méthodes issues de l'ingénierie de la connaissance

Le deuxième axe de contribution est particulièrement intéressant à nos yeux puisque nous avons été amenés à réfléchir sur l'élaboration d'un processus capable d'intégrer et d'exploiter les connaissances d'un expert, dans le but de faciliter l'analyse de règles d'association. Il n'y a pas de solution unique. En effet, la fouille de règles d'association fait intervenir différentes auxquelles ont peut répondre par l'utilisation de techniques adaptées. Un problème nous a pourtant apparu comme « ouvert » : comment éliminer la redondance intrinsèque au domaine d'application et faire ressortir par la même, les règles les plus pertinentes sur le domaine.

L'approche proposée consiste à modéliser une première ébauche des dépendances du domaine, par le biais d'un réseau bayésien. Nous avons ensuite précisé les modalités d'exploitation de ces dépendances : d'une part par la définition d'une mesure d'intérêt permettant le filtrage des règles déjà connues ; d'autre part par l'utilisation des propriétés graphiques du réseau bayésien, à savoir la propriété de d-séparation. Cette propriété a été étudiée dans le but de faire ressortir de l'ensemble des règles présentées à l'expert une décomposition des dépendances présentes dans les règles et prises en compte dans le réseau, de celles qui paraissent surprenantes, car non modélisées. Le but est là aussi de fournir un outil à l'utilisateur qui doit analyser l'ensemble des règles.

Nous avons aussi réfléchi aux interactions de l'expert vis-à-vis des règles. Pour

cela nous avons souhaité développer une interface spécifique regroupant intégrant différentes fonctionnalités que nous avons jugées comme étant indispensables au bon déroulement de l'étape d'étude des résultats : filtrages syntaxiques, tris multiples, tests d'hypothèses (c'est-à-dire la possibilité de valider ou d'infirmer le caractère pertinent d'une règle.

Enfin, pour mémoriser les différentes interactions et interprétations de l'expert, nous avons élaboré un système d'annotations. D'un côté il permet de collecter un ensemble de sous-motifs correspondants aux attributs des règles étudiées. Ces motifs sont catégorisés selon le jugement porté par l'expert (motif connu, non valide, non pertinent, ou pertinent). Ces motifs sont interprétés par le moteur d'affichage des règles, apportant ainsi une aide supplémentaire à l'expert dans la reconnaissance des règles et dans leur analyse. Deuxième intérêt, faciliter la phase de mise à jour du modèle : la collection d'annotations est ainsi envoyée aux experts chargés de faire évoluer le réseau bayésien en fonction des découvertes . Si l'analyse des règles extraites révèlent de nombreux motifs connus, ces motifs doivent être pris en compte dans la structure du réseau afin de faciliter l'analyse des règles pour les itérations ultérieurs (filtrage des motifs connus). Si l'étape d'analyse fait apparaître des motifs pertinents, il faut alors se poser la question suivante : s'agit-il d'un dysfonctionnement du système ? D'une erreur dans les données étudiées ? ou d'une connaissance particulière que l'on souhaite alors intégrer au modèle existant ?

Dans nos travaux nous avons essayé de mettre en avant la constitution d'un cycle vertueux qui tend à converger à la fois vers un modèle de plus en plus riche des dépendances du domaine mais aussi vers la découverte de règles réellement pertinentes pour l'expert du domaine. Une première validation expérimentale a été réalisée sur le cas *Visit Asia*, une deuxième sur le cas d'application industriel.

Cas d'application industriel

Nous avons appliqué notre processus de découverte de connaissances aux données d'interruptions pour le compte d'un grand constructeur aéronautique. En partant d'une ébauche de modèle des dépendances du domaine nous avons élaboré un réseau bayésien dont la structure finale est une bonne représentation des liens existants entre les différentes variables étudiées.

Les précisions successivement apportées au RB ont permis un filtrage de plus en plus efficace dès règles connues, non valides, ou non intéressantes. A la dernière itération de notre processus des règles ayant un fort intérêt par rapport au RB ont effectivement été jugées comme porteuses d'informations pertinentes par l'expert.

Un des avantages constatés expérimentalement par notre approche réside dans la possibilité d'éliminer un grand nombre de règles qui ne présentent pas d'intérêt pour l'expert. De plus, dans nos expériences, aucun faux-négatif n'a été détecté parmi cet

ensemble de règles.

Au final, la découverte de ces dépendances et les modifications qui en ont découlées confortent le principe proposé par notre approche qui est de modéliser, par interaction avec l'expert, les dépendances du domaine et les exploiter pour améliorer le filtrage des règles d'associations extraites et ainsi faciliter leur étude.

Ces résultats soulèvent néanmoins quelques nouvelles interrogations notamment sur la gestion des cas de « faux-positifs ». La solution actuellement apportée est uniquement visuelle. Il faut donc réfléchir au filtrage de ces motifs autrement que par la mesure d'intérêt.

Un deuxième problème a prendre en compte dans des travaux futurs est l'étude de l'impact potentiel que peut avoir la modification de la structure ou des paramètres du RB, sur l'ensemble règles. Il s'agit en quelque sorte de pouvoir contrôler et mesurer les effets de bord provoqués par la mise à jour du RB. Il peut par exemple être intéressant de présenter à l'expert uniquement les règles qui ont été impactées par la dernière modification du modèle.

Perspectives

Ces travaux ont ouvert une perspective intéressante quant à l'étude des liens entre réseau bayésien et règles d'association, Le premier axe consisterait à travailler sur un approfondissement des liens entre réseaux bayésiens et règles d'association. On peut par exemple réfléchir à l'utilisation des règles d'association en complément à l'apprentissage de la structure et des paramètres d'un réseau bayésien ? Le but serait de proposer des modifications du réseau à partir des règles et des annotations.

Plus généralement, nous avons constaté les limites des approches algorithmiques « pures » lorsqu'elles sont appliquées à des cas concrets. Le manque d'approches visant à concilier modèle de connaissances formel et règles d'association apparaît distinctement dans l'état de l'art. Il semble donc crucial d'étudier l'utilisation techniques de fouille au sein d'une approche englobant, à part entière, le facteur humain et les connaissances à disposition sur le domaine d'application. Le but étant de faciliter la découverte de connaissances toujours plus pertinentes sur de grands volumes de données.

Nous avons aussi réalisé que ces deux approches pouvaient se montrer complémentaires : une première ébauche d'une modèle de connaissance initie des découvertes locales sur les données, qui elles mêmes participent à l'élaboration du modèle. Nous pensons que cette démarche est très prometteuse, d'autant plus qu'elle vise à rassembler deux axes de recherches majeurs : l'ingénierie de la connaissance et la fouille de données.

Un autre axe de recherche potentiellement intéressant pourrait être de s'intéresser

au problème de la gestion des évolutions de notre modèle de connaissance, notamment dans un contexte où ce modèle serait partagé entre plusieurs experts. Comment garder une trace des modifications effectuées au fil du temps (pour signaler éventuellement des changements de structure contradictoires) ? Comment différencier et présenter sans ambiguïté, à un instant t du processus, l'ensemble des connaissances implicitement « contenues » dans le réseau bayésien (modélisées a priori, intégrées pour le filtrage de motifs, ou encore les connaissances nouvelles) ?

Annexe A

Présentation de l'application

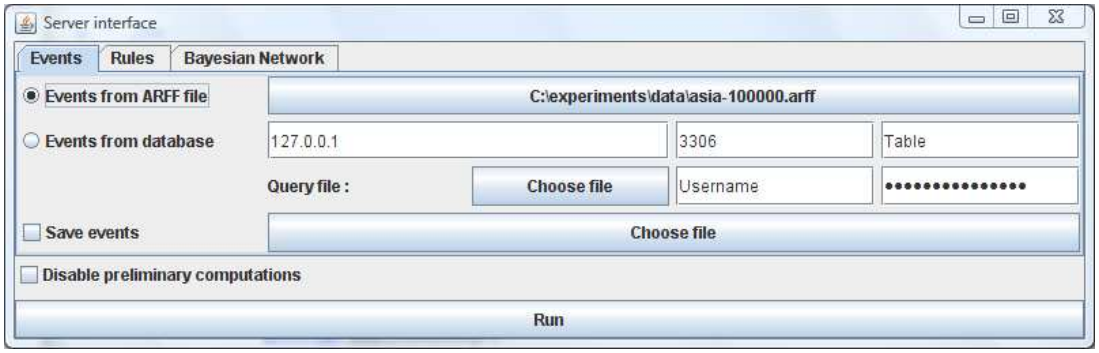

FIG. A.1 – Interface de configuration du serveur : onglet permettant la configuration des sources de données.

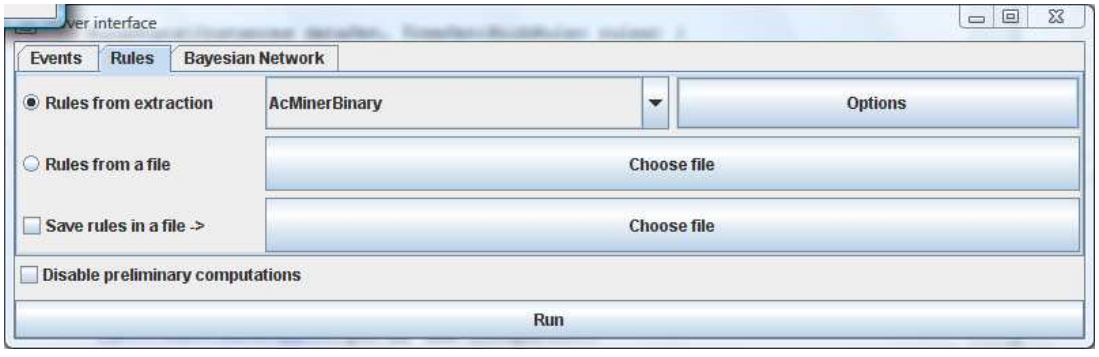

FIG. A.2 – Interface de configuration du serveur : onglet de configuration de l'algorithme d'extraction.

[]

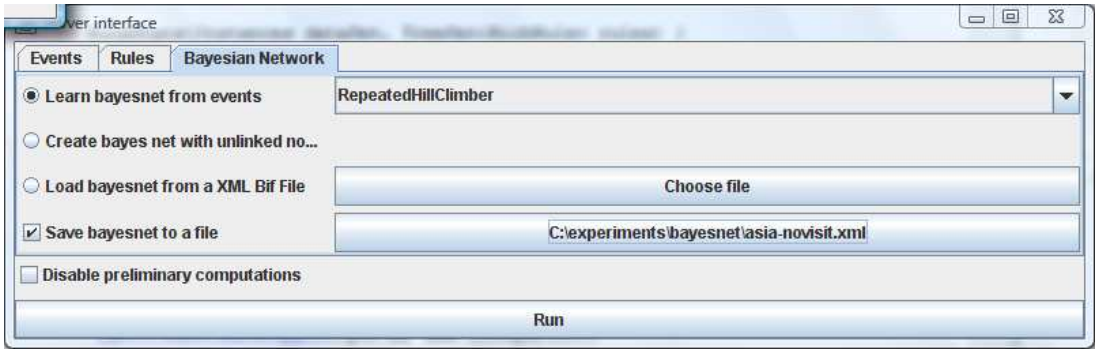

FIG. A.3 – Interface de configuration du serveur : onglet permettant la configuration du réseau bayésien initial.

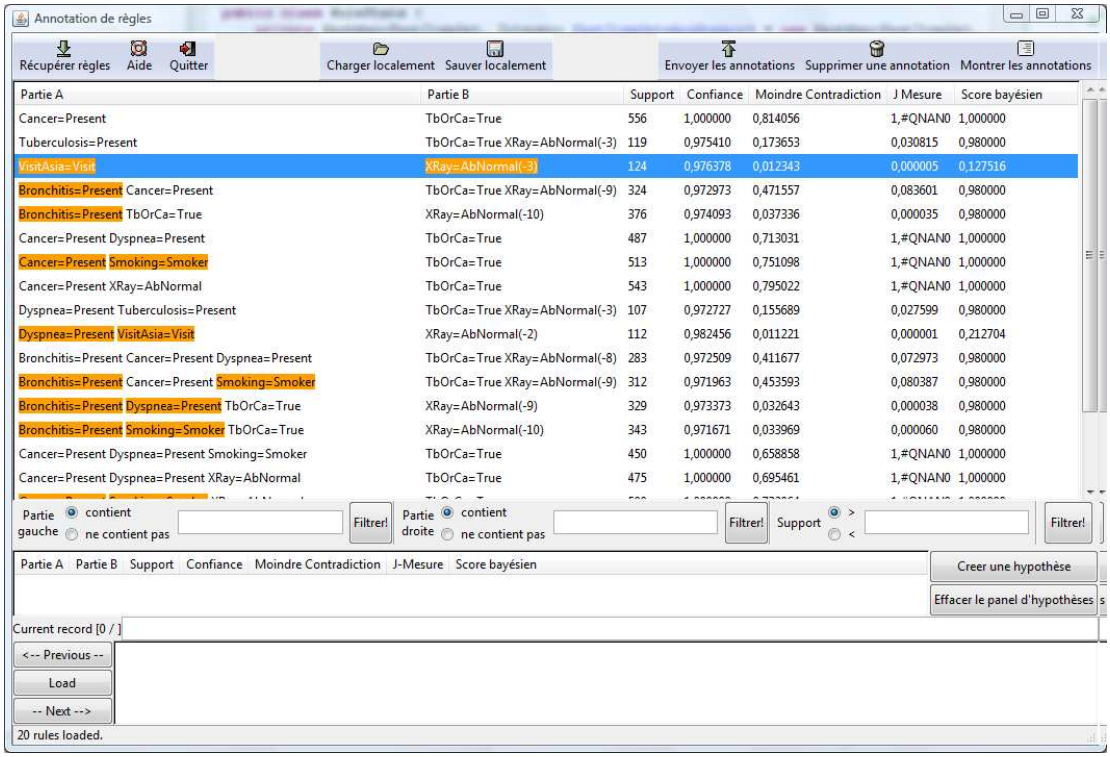

FIG. A.4 – Interface d'analyse des règles d'association.

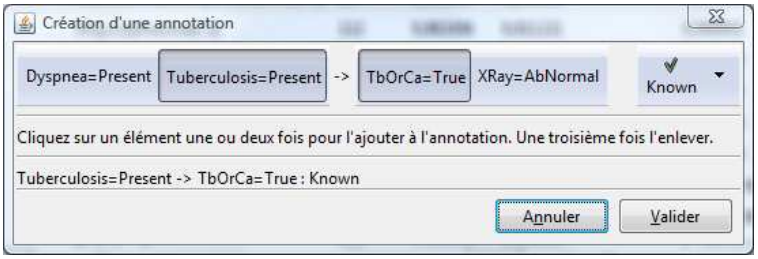

FIG. A.5 – Annotation de règles d'association.

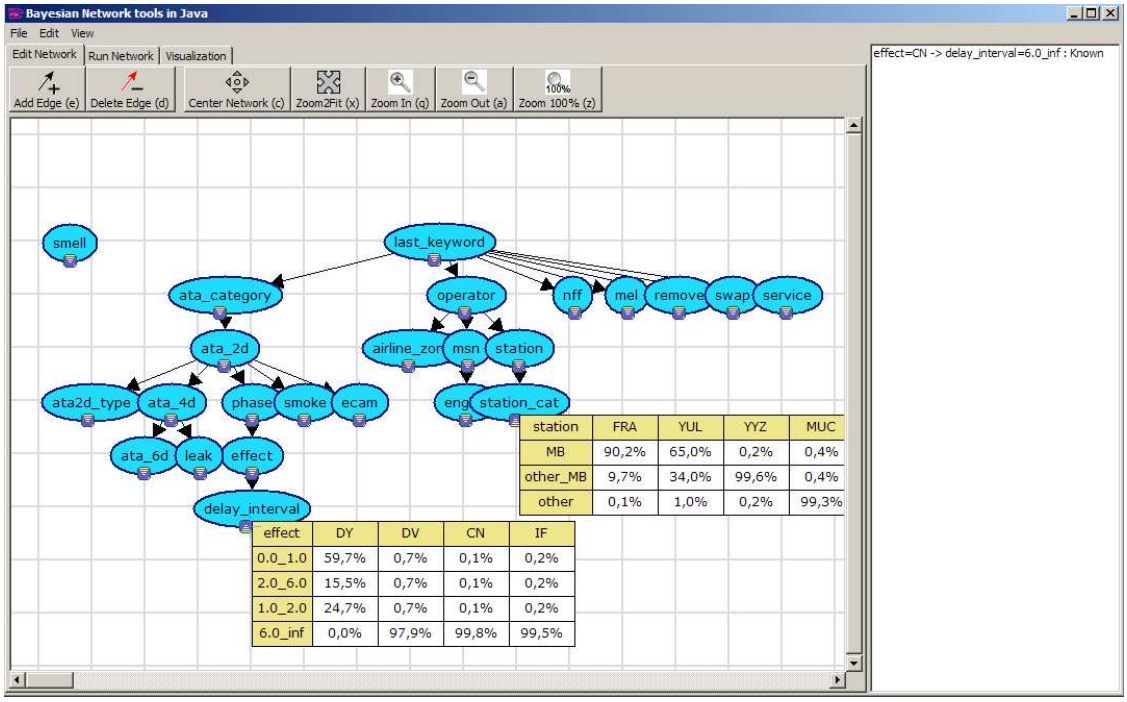

FIG. A.6 – Mise à jour du réseau bayésien à partir des annotations.

Bibliographie

[AA07] D.J. Newman A. Asuncion. UCI machine learning repository, 2007.

[AAP00] Ramesh C. Agarwal, Charu C. Aggarwal, and V. V. V. Prasad. Depth first generation of long patterns. In Proceedings of the 6th international conference on knowledge discovery and data mining (KDD'00), 2000.

[AAP01] Ramesh C. Agarwal, Charu C. Aggarwal, and V. V. V. Prasad. A tree projection algorithm for generation of frequent item sets. Journal of Parallel and Distributed Computing, 61(3) :350–371, 2001.

[AIS93] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In Peter Buneman and Sushil Jajodia, editors, Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pages 207–216, Washington, D.C., 26–28 1993.

[AK01] Jérôme Azé and Yves Kodratoff. Evaluation de la résistance au bruit de quelques mesures d'extraction de règles d'assocation. Extraction des connaissances et apprentissage, 1(4) :143–154, 2001.

[AL99] Yonatan Aumann and Yehuda Lindell. A statistical theory for quantitative association rules. In KDD '99 : Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 261–270, New York, NY, USA, 1999. ACM Press.

[AMS+96] Rakesh Agrawal, Hiekki Mannila, Ramakrishnan Srikant, Hannu Toivonen, and A. Inkeri Verkamo. Fast discovery of association rules, chapter 12, pages 307–328. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996.

[AS94] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, Proceedings of the 1994 VLDB International Conference on Very Large Data Bases, pages 487–499. Morgan Kaufmann, 12–15 1994.

[Azé03] Jérôme Azé. Extraction de connaissances à partir de données numériques et textuelles. thèse de doctorat, Université Paris-Sud, december 2003.

[BAG00] R.J. Bayardo, Rakesh Agrawal, and D. Gunopulos. Constraint-based rule mining in large, dense databases. Data mining and knowledge discovery, 4 :30–59, 2000.

[Bay98] R.J. Bayardo. Efficiently mining long patterns from databases. In Proceedings of the SIGMOD conference, pages 85–93, 1998.

[BB00] Jean-François Boulicaut and Artur Bykowski. Frequent closures as a concise representation for binary data mining. In Proceedings of the 2000 PAKDD Pacific-Asia Conference on Knowledge Discovery and Data Mining, volume 1805 of LNAI, pages 62–73, Kyoto, JP, April 2000. Springer-Verlag.

[BBJ+02] Céline Becquet, Sylvain Blachon, Baptiste Jeudy, Jean-François Boulicaut, and Olivier Gandrillon. Strong association rule mining for large gene expression data analysis : a case study on human SAGE data. Genome Biology, 12 :–, 2002. See http ://genomebiology.com/2002/3/12/research/0067.

[BBR00] Jean-Francois Boulicaut, Artur Bykowski, and Christophe Rigotti. Approximation of frequency queries by means of free-sets. In Proceedings of the 2000 PKDD European Conference on Principles and Practice of Knowledge Discovery in Databases, pages 75–85, 2000.

[BBR03] Jean-François Boulicaut, Artur Bykowski, and Christophe Rigotti. Free-sets : A condensed representation of boolean data for the approximation of frequency queries. Data Min. Knowl. Discov., 7(1) :5–22, 2003.

[BGH+02] R. Barco, R. Guerrero, G. Hylander, L. Nielsen, M. Partanen, and S. Patel. Automated troubleshooting of mobile networks using bayesian networks. In IASTED International Conference on Communication Systems and Networks, Malaga, Spain, 2002.

[BGS+00] Tom Brijs, Bart Goethals, Gilbert Swinnen, Koen Vanhoof, and Geert Wets. A data mining framework for optimal product selection in retail supermarket data : the generalized PROFSET model. In R. Ramakrishnan, S. Stolfo, R. Bayardo, and I. Parsa, editors, Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, August 20-23, 2000, Boston, MA, USA. ACM, 2000, pages 300–304, 2000.

[BKGG04] Julien Blanchard, Pascale Kuntz, Fabrice Guillet, and Régis Gras. Mesure de qualité des règles d'association par l'intensité d'implication entropique. Revue Nationale des Technologies de l'Information, E(1) :33–44, 2004.

[BKM99] Jean-Francois Boulicaut, Mika Klemettinen, and Heikki Mannila. Modeling kdd processes within the inductive database framework. In DaWaK '99 : Proceedings of the First International Conference on Data

Warehousing and Knowledge Discovery, pages 293–302, London, UK, 1999. Springer-Verlag.

[BMM03] Bernadette Bouchon-Meunier and Christophe Marsala. Logique floue, principes, aide à la décision. Editions Hermes, 2003.

[BMUT97] Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman, and Shalom Tsur. Dynamic itemset counting and implication rules for market basket data. In Joan Peckham, editor, SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, May 13-15, 1997, Tucson, Arizona, USA, pages 255–264. ACM Press, 05 1997.

[BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1–7) :107–117, 1998.

[BRM02] M. Brodie, I. Rish, and S. Ma. ntelligent probing : A cost-effective approach to fault diagnosis in computer networks. IBM Systems Journal, 41 :–, 2002.

[BTP+00] Yves Bastide, Rafik Taouil, Nicolas Pasquier, Gerd Stumme, and Lotfi Lakhal. Mining frequent patterns with counting inference. SIGKDD Explorations, 2(2) :66–75, Décembre 2000.

[BTP+02] Yves Bastide, Rafik Taouil, Nicolas Pasquier, Gerd Stumme, and Lotfi Lakhal. Pascal : un algorithme d'extraction de motifs fréquents. Techniques et Science Informatiques, 21(1) :65–95, 2002.

[CB02] Bruno Crémilleux and Jean-François Boulicaut. Simplest rules characterizing classes generated by δ-free sets. In Springer, editor, Proceedings of the twenty-second Annual International Conference Knowledge Based Systems and Applied Artificial Intelligence (ES'02), pages 33–46, December 2002.

[CCK+00] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth. Crisp-dm 1.0 step-by-step data mining guide. CRISP-DM Consortium, 2000.

[CDLS03] Robert G. Cowell, A.Philip Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter. Probabilistic Networks and Expert Systems. Springer, 2003.

[CGK+02] Jie Cheng, Russell Greiner, Jonathan Kelly, David Bell, and Weiru Liu. Learning bayesian networks from data : An information-theory based approach. Artificial Intelligence, 137 :309–347, 2002.

[CHM04] David Maxwell Chickering, David Heckerman, and Christopher Meek. Large-sample learning of bayesian networks is np-hard. Journal of Machine Learning Research, 5 :1287–1330, 2004.

[Chr75] Nicos Christofides. Graph theory : An algorithmic approach (computer science and applied mathematics). 1975.

[Coo88] G. F. Cooper. Probabilistic inference using belief networks is np-hard. Technical report, Knowledge Systems Laboratory, 1988.

[DD00] Marek J. Druzdzel and F. Diez. Criteria for combining knowledge from different sources in probabilistic networks, 2000.

[Dec99] Rina Dechter. Bucket elimination : A unifying framework for reasoning. Artificial Intelligence, 113(1-2) :41–85, 1999.

[DG00] Marek J. Druzdzel and Linda C. Van Der Gaag. Building probabilistic networks : 'where do the numbers come from ?' guest editors' introduction. IEEE Transactions on Knowledge and Data Engineering, 12(4) :481–486, 2000.

[DL99] Guozhu Dong and Jinyan Li. Efficient mining of emerging patterns : discovering trends and differences. In KDD '99 : Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 43–52, New York, NY, USA, 1999. ACM Press.

[DLR77] Arthur Dempster, Nan Laird, and Donald Rubin. Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society, Series B, 39(1) :1–38, 1977.

[DP01] William DuMouchel and Daryl Pregibon. Empirical bayes screening for multi-item associations. In KDD '01 : Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pages 67–76, New York, NY, USA, 2001. ACM Press.

[FDBM06] Clément Fauré, Sylvie Delprat, Jean-François Boulicaut, and Alain Mille. Iterative bayesian network implementation by using annotated association rules. In EKAW, pages 326–333, 2006.

[FDMB06a] Clément Fauré, Sylvie Delprat, Alain Mille, and Jean-François Boulicaut. Utilisation des réseaux bayésiens dans le cadre de l extraction de règles d association. In Actes de la conférence EGC'2006 pour l Extraction et la Gestion des connaissances, 2006.

[FDMB06b] Clément Fauré, Sylvie Delprat, Alain Mille, and Jean-François Boulicaut. Utilisation des réseaux bayésiens dans le cadre de l'extraction de règles d'association. In EGC, pages 569–580, 2006.

[FL04] Olivier François and Philippe Leray. Etude comparative d'algorithmes d'apprentissage de structure dans les réseaux bayésiens. Journal électronique d'intelligence artificielle, 5(39) :1–19, 2004.

[FMMT96] Takeshi Fukuda, Yasuhido Morimoto, Shinichi Morishita, and Takeshi Tokuyama. Mining optimized association rules for numeric attributes. In PODS '96 : Proceedings of the fifteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems, pages 182–191, New York, NY, USA, 1996. ACM Press.

[FP96] Tom Fawcett and Foster Provost. Combining data mining and machine learning for effective user profiling. In Simoudis, Han, and Fayyad, editors, Proceedings on the Second International Conference on Knowledge Discovery and Data Mining, pages 8–13, Menlo Park, CA, 1996. AAAI Press.

[FP97] Tom Fawcett and Foster J. Provost. Adaptive fraud detection. Data Mining and Knowledge Discovery, 1(3) :291–316, 1997.

[FPSM92] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus. Knowledge discovery in databases - an overview. Ai Magazine, 13 :57–70, 1992.

[GCB+04] Régis Gras, Raphaël Couturier, Maurice Bernadet, Julien Blanchard, Henri Briand, Fabrice Guillet, Pascale Kuntz, Rémi Lehn, and Philippe Peter. Quelques critères pour une mesure de qualité de règles d'association - un exemple : l'intensité d'implication. Revue des Nouvelles Technologies de l'Information (RNTI-E), 1, 2004.

[GRS96] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. Markov Chain Monte Carlo In Practice. Chapman & Hall, 1996.

[GZ03] Bart Goethals and Mohammed J. Zaki. Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations FIMI 2003. Melbourne, USA, 2003.

[HC99] John D. Holt and Soon M. Chung. Efficient mining of association rules in text databases. In CIKM '99 : Proceedings of the eighth international conference on Information and knowledge management, pages 234–242, New York, NY, USA, 1999. ACM Press.

[HCC02] Emmanuel Hugues, Eric Charpentier, and André Cabarbaye. Applicatin of markov processes to predict aircraft operational reliability. In Proceedings of the 3rd European systems engineering conference, 2002.

[Hec95] D. Heckerman. A tutorial on learning with bayesian networks. Technical report, Microsoft Research, 1995.

[Hec97] David Heckerman. Bayesian networks for data mining. Data Mining and Knowledge Discovery, 1(1) :79–119, 1997.

[HF95] Jiawei Han and Yongjian Fu. Discovery of multiple-level association rules from large databases. In VLDB '95 : Proceedings of the 21th International Conference on Very Large Data Bases, pages 420–431, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[HG94] Peter E. Hart and Jamey Graham. Query-free information retrieval. In Conference on Cooperative Information Systems, pages 36–46, 1994.

[HGC94] David Heckerman, Dan Geiger, and David Maxwell Chickering. Learning bayesian networks : The combination of knowledge and statistical data. In KDD Workshop, pages 85–96, 1994.

[HGN00] Jochen Hipp, Ulrich Güntzer, and Gholamreza Nakhaeizadeh. Algorithms for association rule mining - a general survey and comparison. SIGKDD Explor. Newsl., 2(1) :58–64, 2000.

[HH99] Robert J. Hilderman and Howard J. Hamilton. Knowledge discovery and interestingness measures : a survey. Technical report, Department of Computer Science, University of Regina, 1999.

[HK00] Jiawei Han and Micheline Kamber. Data mining : concepts and techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000.

[HKMT95] Marcel Holsheimer, Martin L. Kersten, Heikki Mannila, and Hannu Toivonen. A perspective on databases and data mining. In 128, page 10. Centrum voor Wiskunde en Informatica (CWI), ISSN 0169-118X, 30 1995.

[HMS01] David Hand, Heikki Mannila, and Padhraic Smyth. Principles of data mining. The MIT Press, 2001.

[HMWG98] Jochen Hipp, Andreas Myka, Rudiger Wirth, and Ulrich Guntzer. A new algorithm for faster mining of generalized association rules. pages 74–82, 1998.

[IM96] Tomasz Imielinski and Heikki Mannila. A database perspective on knowledge discovery. Communications of the ACM, 39(11) :58–64, 1996.

[IV99] Tomasz Imieliński and Aashu Virmani. Msql : A query language for database mining. Data Min. Knowl. Discov., 3(4) :373–408, 1999.

[Jeu02] Baptiste Jeudy. Optimisation de requêtes inductives : application à l'extraction sous contraintes de règles d'association. PhD thesis, Université Lyon I, INSA de Lyon, december 2002.

[JGJS99] Michael I. Jordan, Zoubin Ghahramani, Tommi Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. Machine Learning, 37(2) :183–233, 1999.

[JJJ96] Finn V. Jensen, F.V. V. Jensen, and F. V. Jensen. Introduction to Bayesian Networks. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996.

[Jor98] M. I. Jordan. Learning in Graphical Models. MIT Press, 1998.

[JrK+] Finn V. Jensen, Uffe Kjærulff, Brian Kristiansen, Claus Skaanning Helge Langseth, Jiri Vomlel, and Marta Vomlelova. The sasco methodology for troubleshooting complex systems.

[JS02] S. Jaroszewicz and D. A. Simovici. Pruning redundant association rules using maximum entropy principle. In Advances in Knowledge Discovery and Data Mining, 6th Pacific-Asia Conference, PAKDD'02, pages 135–147, Taipei, Taiwan, May 2002.

[JS04] Szymon Jaroszewicz and Dan A. Simovici. Interestingness of frequent itemsets using bayesian networks as background knowledge. In Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 178–186, New York, NY, USA, 2004. ACM Press.

[JS05] Szymon Jaroszewicz and Tobias Scheffer. Fast discovery of unexpected patterns in data, relative to a bayesian network. In Proceedings of the 2005 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2005. ACM Press.

[Kle99] Mika Klemettinen. A Knowledge Discovery Methodology for Telecommunication Network Alarm Databases. PhD thesis, University of Helsinki, 1999.

[KMR+94] Mika Klemettinen, Heikki Mannila, Pirjo Ronkainen, Hannu Toivonen, and A. Inkeri Verkamo. Finding interesting rules from large sets of discovered association rules. In CIKM '94 : Proceedings of the third international conference on Information and knowledge management, pages 401–407, New York, NY, USA, 1994. ACM Press.

[Kra98] P. Krause. Learning probabilistic networks, 1998.

[LAK+01] R. D. Lawrence, G. S. Almasi, V. Kotlyar, M. S. Viveros, and S. S. Duri. Personalization of supermarket product recommendations. Data Min. Knowl. Discov., 5(1-2) :11–32, 2001.

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

[LK98] D. Lin and Z.M. Kedem. Pincer-search : a new algorithm for discovering the maximum frequent set. In Proceedings of the EBDT conference, pages 105–119, 1998.

[LMV+04] Philippe Lenca, Patrick Meyer, Benoît Vaillant, Picouet P., and Stéphane Lallich. Evaluation et analyse multi-critères des mesures de qualité des règles d'association. Revue des Nouvelles Technologies de l'Information (RNTI-E), 1 :219–246, 2004.

[LPT04] Stéphane Lallich, Elie Prudhomme, and Olivier Teytaud. Contrôle du risque multiple pour la sélection de règles d'association significatives. In Actes de la 4e Conférence EGC Extraction et Gestion des Connaissances, volume 2 of Revue des Nouvelles Technologies de l'information (RNTI-E-2), pages 193–217, 2004.

[LS88] Steffen Lauritzen and David Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. Journal of the Royal Statistical Society, Series B, 50(2) :157–224, 1988.

[LSM00] Wenke Lee, Salvatore J. Stolfo, and Kui W. Mok. Adaptive intrusion detection : A data mining approach. Artificial Intelligence Review, 14(6) :533–567, 2000.

[LT03] S. Lallich and O. Teytaud. Evaluation et validation de l'intérêt des règles d'association. Revue des Nouvelles Technologies de l'Information, n° 2 :–, 2003.

[LT04] Stéphane Lallich and Olivier Teytaud. Évaluation et validation de l'intérêt des règles d'association. Revue des Nouvelles Technologies de l'Information (RNTI-E), 1 :193–217, 2004.

[MAA05] Ahmed Metwally, Divyakant Agrawal, and Amr El Abbadi. Using association rules for fraud detection in web advertising networks. In VLDB '05 : Proceedings of the 31st international conference on Very large data bases, pages 169–180. VLDB Endowment, 2005.

[MPC96] Rosa Meo, Giuseppe Psaila, and Stefano Ceri. A new sql-like operator for mining association rules. In VLDB '96 : Proceedings of the 22th International Conference on Very Large Data Bases, pages 122–133, San Francisco, CA, USA, 1996. Morgan Kaufmann Publishers Inc.

[MPC98] Rosa Meo, Giuseppe Psaila, and Stefano Ceri. An extension to sql for mining association rules. Data Mining and Knowledge Discovery, 2(2) :195–224, 1998.

[MT96] Heikki Mannila and Hannu Toivonen. Multiple uses of frequent sets and condensed representations. In Proceedings of the 1996 KDD International Conference on Knowledge Discovery and Data Mining, pages 189–194, 1996.

[MT97] Heikki Mannila and Hannu Toivonen. Levelwise search and borders of theories in knowledge discovery, volume 3, chapter 1, pages 241–258. KluwerAcademic Publishers, 1997.

[MTV94] Heikki Mannila, Hannu Toivonen, and A.I. Verkamo. Efficient algorithms for discovering association rules. In Proceedings of the AAAI Workshop on Knowledge Discovery in Databases, pages 181–192, 1994.

[Nea93] R. M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, University of Toronto, 1993.

[NH98] R. Neal and G. Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, Learning in Graphical Models. Kluwer, 1998.

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

[NWL+04] Patrick Naïm, Pierre-Henri Wuilleminn, Philippe Leray, Olivier Pourret, and Anna Becker. Réseaux bayésiens. Eyrolles, 05 2004.

[ODW00] Agnieska Onisko, Marek Druzdzel, and Hanna Wasyluk. Learning bayesian network parameters from small data sets : Application of noisy-or gates, 2000.

[Pas00] Nicolas Pasquier. Data mining : algorithmes d'extraction et de réduction des règles d'association dans les bases de données. thèse de doctorat, Université Clermont-Ferrand II, LIMOS, Complexe scientifique des Céseaux, F-63177 Aubière cedex, France, december 2000.

[PBTL98] Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal. Pruning closed itemset lattices for association rules. In Proceedings of the BDA Conference, pages 177–196, 1998.

[PBTL99a] Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal. Discovering frequent closed item for association rules. In Catriel Beeri and Peter Buneman, editors, Proceedings of the 7th international conference on knowledge discovery and data mining, pages 398–416, Jerusalem, Israël, January 1999. Berlin : Springer-Verlag.

[PBTL99b] Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal. Efficient mining of association rules using closed itemset lattices. Information Systems, 24(1) :25–46, January 1999.

[Pea88] Judea Pearl. Probabilistic reasoning in intelligent systems : networks of plausible inference. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[PHM00] Jian Pei, Jiawei Han, and Runuying Mao. Closet an efficient algorithm for mining frequent closed itemsets. In Dimitrios Gunopulos and Rajeev Rastogi, editors, Proceedings of the ACM SIGMOD workshop on research issues in data mining and knowledge discovery (DMKD'00)., 2000.

[PKS+03] C. Potter, S. Klooster, M. Steinbach, P. Tan, V. Kumar, S. Shekhar, R. Nemani, and R. Myneni. Global teleconnections of ocean climate to terrestrial carbon flux. Journal of Geophysical Research - Atmospheres, 108(D17), September 2003.

[Pol66] Michael Polanyi. Tacit Dimension. Routledge & Kegan Paul Ltd, London, 1966.

[PPMH94] M. Pradhan, G. Provan, B. Middleton, and M. Henrion. Knowledge engineering for large belief networks. In Uncertainty in AI, Proceedings of the Tenth Conference, 1994.

[PT98] Balaji Padmanabhan and Alexander Tuzhilin. A belief-driven method for discovering unexpected patterns. In Proceedings of the 1998 KDD International Conference on Knowledge Discovery and Data Mining, pages 94–100, 1998.

[PT00] Balaji Padmanabhan and Alexander Tuzhilin. Small is beautiful : discovering the minimal set of unexpected patterns. In Proceedings of the

2000 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 54–63, New York, NY, USA, 2000. ACM Press.

[PT06] Balaji Padmanabhan and Alexander Tuzhilin. On characterization and discovery of minimal unexpected patterns in rule discovery. IEEE Trans. Knowl. Data Eng., 18(2) :202–216, 2006.

[PTB+05] Nicolas Pasquier, Rafik Taouil, Yves Bastide, Gerd Stumme, and Lotfi Lakhal. Generating a condensed representation for association rules. Journal of Intelligent Information Systems, 24 :1 :29–60, 2005.

[PuG+02] S. Populaire, T. Denœux, A. Guilikeng, P. Gineste, and J. Blanc. Fusion of expert knowledge with data using belief functions : a case study in wastewater treatment. In Proceedings of the 5th international Conference on Information Fusion, 2002.

[Rae00] Luc De Raedt. A logical database mining query language. In ILP '00 : Proceedings of the 10th International Conference on Inductive Logic Programming, pages 78–92, London, UK, 2000. Springer-Verlag.

[Ren01] Silja Renooij. Probability elicitation for belief networks : issues to consider. Cambridge University Press, 16 :255–269, 2001.

[RJBA99] Jr. Roberto J. Bayardo and Rakesh Agrawal. Mining the most interesting rules. In Proceedings of the 1999 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 145–154, New York, NY, USA, 1999. ACM Press.

[Rob96] Christian P. Robert. The Bayesian Choice : A Decision-Theoretic Motivation. Springer-Verlag, 1996.

[rul94] Uffe Kjærulff. Reduction of computational complexity in bayesian networks through removal of weak dependences. In Proceedings of the 10th conference on uncertainty in Artificial Intelligence, pages 374–382. Association for Uncertainty in Artificial Intelligence, Morgann Kaufmann Publishers, 1994.

[RW99] S. Renooij and C. Witteman. Talking probabilities : communicating probabilistic information with words and numbers. International Journal of Approximate Reasoning, 22 :169–194, 1999.

[SA96] Ramakrishnan Srikant and Rakesh Agrawal. Mining quantitative association rules in large relational tables. In SIGMOD '96 : Proceedings of the 1996 ACM SIGMOD international conference on Management of data, pages 1–12, New York, NY, USA, 1996. ACM Press.

[SCH05] Wang Shitong, Korris F. L. Chung, and Shen Hongbin. Fuzzy taxonomy, quantitative database and mining generalized association rules. Intell. Data Anal., 9(2) :207–217, 2005.

[SG92] P. Smyth and R. M. Goodman. An information theoretic approach to rule induction from databases. IEEE Transactions on Knowledge and Data Engineering, 4(4) :301–316, 1992.

[Sme00] P. Smets. Data fusion in the transferable belief model. In in Proceedings of the Third International Conference on Information Fusion, 2000.

[SON95] Ashoka Savasere, Edward Omiecinski, and Shamkant B. Navathe. An efficient algorithm for mining association rules in large databases. In Proceedings of the 21th International Conference on Very Large Data Bases, 1995.

[SSO+97] K. Satou, G. Shibayama, T. Ono, Y. Yamamura, E. Furuichi, S. Kuhara, and T. Takagi. Finding association rules on heterogeneous genome data. In Proceedings of the Pacific Symposium on Biocomputing, pages 397–408, 1997.

[ST96] Avi Silberschatz and Alexander Tuzhilin. What makes patterns interesting in knowledge discovery systems. IEEE Transactions on Knowledge and Data Engineering, 8(6) :970–974, 1996.

[STVN04] Ansaf Salleb, Teddy Turmeaux, Christel Vrain, and Cyril Nortet. Mining quantitative association rules in a atherosclerosis dataset. In PKDD Discovery Challenge 2004 (co-located with the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases), pages 98–103, september 2004.

[TMH01] Bo Thiesson, Christopher Meek, and David Heckerman. Accelerating em for large databases. Mach. Learn., 45(3) :279–299, 2001.

[Toi96] Hannu Toivonen. Sampling large databases for association rules. In VLDB '96 : Proceedings of the 22th International Conference on Very Large Data Bases, pages 134–145, San Francisco, CA, USA, 1996. Morgan Kaufmann Publishers Inc.

[TPBL00] R. Taouil, N. Pasquier, Y. Bastide, and L. Lakhal. Mining bases for association rules using closed sets. In Proceedings of the ICDE conference, page 307, 2000.

[vdGRW+02] L. van der Gaag, S. Renooij, C. Witteman, B. Aleman, and B. Taal. Probabilities for a probabilistic network : A case-study in oesophageal cancer, 2002.

[WF05] Ian H. Witten and Eibe Frank. Data Mining : Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco, 2nd edition edition, 2005.

[XHD+05] Hui Xiong, Xiaofeng He, Chris Ding, Ya Zhang, Vipin Kumar, and Stephen R. Holbrook. Identification of functional modules in protein complexes via hyperclique pattern discovery. In Proceedings of Pacific Symposium on Biocomputing, volume 10, 2005.

[Zak00] Mohammed Javeed Zaki. Generating non-redundant association rules. In Proceedings of the KDD Conference (FIXME), 2000.

[ZH02] Mohammed Javeed Zaki and Ching-Jui Hsiao. Charm : an efficient algorithm for closed itemset mining. In Robert Grossman, Jiawei Han, Vipin Kumar, Heikki Mannila, and Rajeev Motwani, editors, Proceedings of the 2nd international SIAM conference on data mining (SDM'02), 2002.

[ZO98] Mohammed Javeed Zaki and M. Ogihara. Theoritical foundations of association rules. In Proceedings of the DMKD Workshop on Research Issues in Data Mining and Knowledge Discovery, pages 7 :1–7 :8, 1998.

[ZPOL97] Mohammed Javeed Zaki, S. Parthasarathy, M. Ogihara, and W. Li. New algorithms for fast discovery of association rules. In Proceedings of the KDD conference, pages 283–286, 1997.