# Construction itérative d'un modèle de connaissance par l'exploitation de règles d'association

Clément Fauré, Sylvie Delprat, Jean-François Boulicaut

# Construction itérative d'un modèle de connaissance par l'exploitation de règles d'association

Clément Fauré[1,2], Sylvie Delprat[1], Alain Mille[3], Jean-François Boulicaut[2]

[1] EADS CCR, Département Systèmes Apprenants, Centreda 1, F-31700 Blagnac
{clement.faure, sylvie.delprat}@eads.net

[2] LIRIS UMR 5205, INSA Lyon, Bâtiment Blaise Pascal, F-69621 Villeurbanne cedex
[3] LIRIS UMR 5205, Université Lyon 1, Nautibus, F-69622 Villeurbanne cedex
{amille, jboulica}@liris.cnrs.fr

**Résumé** Nous nous intéressons à la construction itérative d'un modèle de la connaissance experte par l'exploitation de règles descriptives telles que les règles d'associations. Nous avons montré que, lorsqu'il est disponible, un modèle de type réseau bayésien facilite la présentation de règles d'association pertinentes. Nous étudions maintenant les possibilités pour l'expert d'annoter ces règles pour d'une part converger vers un modèle de connaissance « augmenté » mais aussi vers la découverte de règles toujours plus intéressantes. Nous considérons les différents problèmes liés à la construction, l'exploitation et la mise à jour de tels modèles ainsi que nos éléments de solution. Concrètement, cette étude s'appuie sur un cas d'application à l'analyse des interruptions opérationnelles dans l'industrie aéronautique.

**Mots-clés** : Ingénierie des connaissances, réseaux baysésiens, règles d'association.

## 1 Introduction

L'un des objectifs de l'extraction de connaissances à partir de données consiste à fournir des énoncés valides et utiles aux utilisateurs propriétaires de ces données. L'utilité de ces énoncés est d'autant plus grande qu'ils décrivent une réalité du domaine qui n'a pas encore été explicitée jusqu'ici, autrement dit, une nouvelle connaissance.

Nous nous intéressons à la découverte de connaissances au moyen de règles descriptives comme les règles d'association [1]. Notre hypothèse de travail est que les règles dites « intéressantes » sont celles qui non seulement satisfont certaines contraintes sur des mesures d'intérêt objectives (e.g., confiance, fréquence) mais aussi qui sortent du cadre des connaissances déjà connues de l'utilisateur. On peut alors imaginer de nouvelles mesures d'intérêt qui quantifient le degré de nouveauté au regard d'un modèle de la connaissance experte. Ainsi, [13] a revisité la notion de fréquence significative lorsqu'un réseau bayésien spécifie les distributions "attendues". Dans [9], nous sommes parti 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. Il nous faut donc étudier précisément les interactions entre l'expert d'une part et le réseau bayésien qui modélise une partie de sa connaissance d'autre part. Quelles sont les principales difficultés rencontrées lors de la construction et de l'exploitation d'un réseau bayésien dans ces contextes de fouille de données ? Quelles sont les solutions existantes ? Quelles pistes pouvons nous envisager pour résoudre les problèmes ouverts ? C'est ce que nous proposons de décrire dans la suite de cet article en nous appuyant sur un cas d'application pratique : la fouille de données d'interruptions opérationnelles en aéronautique.

La suite de l'article est organisée de la façon suivante : la section 2 présente l'approche méthodologique envisagée pour faciliter la découverte de connaissances à base de règles d'association. La section 3 aborde les principales difficultés liées à la mise en place et à l'exploitation d'un réseau bayésien dans notre contexte. Les différentes étapes du processus de fouille sont illustrées sur un cas d'application. La dernière section est une brève conclusion sur les intérêts et les limites de notre approche.

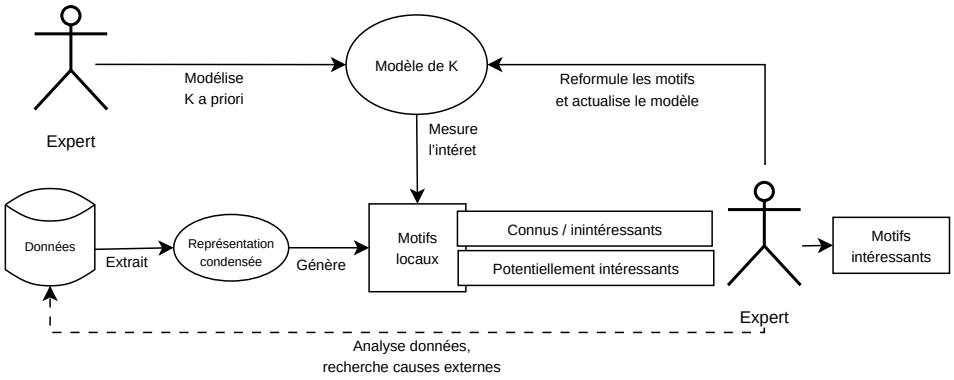


**FIG. 1** – *Aperçu du processus de découverte de connaissances étudié*

# 2 Modélisation et exploitation de la connaissance pour la fouille de données

## 2.1 Approche envisagée

L'approche proposée initialement dans [9] comporte quatre points principaux :

1. Explicitation et modélisation des connaissances a priori de l'expert.
2. Génération d'un ensemble concis de règles d'association.
3. Utilisation du modèle de connaissance pour faciliter la lecture des règles d'association extraites.
4. Prise en compte des résultats de l'analyse (modification de la structure ou des paramètres du réseau bayésien, modification de la base de données).

La figure 1 reprend les différentes étapes de ce processus de découverte de connaissances et en montre une vue globale. En particulier, ce processus repose sur l'extraction d'une collection concise de règles d'association, puis sur la construction et l'exploitation de certaines connaissances de l'expert dans le but de filtrer les règles d'association ininteressantes. Dans la section 3, nous précisons les problèmes posés via des exemples sur un cas d'application pratique.

## 2.2 Extraction de règles

Cette phase concerne l'utilisation d'un algorithme d'extractions de motifs, en l'occurrence des règles d'association. Les algorithmes de type Apriori [2] permettent d'extraire toutes les règles d'association au dessus de seuils de de fréquence et de confiance spécifiés par l'utilisateur. Un premier reproche classique vis à vis de ces algorithmes est qu'ils ne sont pas utilisables sur des données denses et/ou fortement corrélées, tout du moins pour les seuils de fréquences qui paraissent pertinents aux experts. Un second problème vient du fait que toutes les règles qui satisfont les contraintes de fréquence et de confiance sont extraites, ce qui peut déboucher couramment sur des centaines de milliers de règles. De nombreuses règles présentent alors des informations redondantes et, pour les experts, les indispensables tâches d'interprétation s'avèrent fastidieuses ou même impossibles. La question de la redondance de ces collections, quel que soit le domaine d'application, a été très étudiée (voir les nombreuses propositions de couvertures de collections de règles, [11] par exemple). Par contre, au delà de stratégies simples (exploitation de taxonomies sur les attributs en cours d'extraction, stratégies à base de "templates" en post-traitement), il y a encore peu de travaux sur l'élimination automatique de motifs redondants au regard de connaissances déjà acquises par l'expert.

Pour résoudre le premier problème, nous utilisons un algorithme [4] capable d'extraire une représentation condensée des ensembles fréquents, les ensembles dits $\delta$-libres fréquents. Cet algorithme permet également, en calculant la $\delta$-fermeture de tels ensembles, de produire une collection concise de règles d'association à forte confiance appelées règles $\delta$-fortes. En effet, le paramètre $\delta$ détermine le nombre d'exceptions tolérées pour les règles et sa valeur est supposée être petite au regard du seuil de fréquence utilisé. N. Pasquier [17] a d'ailleurs étudié les propriétés de ces collections lorsque $\delta = 0$.

Pour remédier au second problème, la proposition que nous avons faite dans [9] consiste à intégrer la connaissance de l'expert 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 l'émergence de règles plus intéressantes. Plutôt que d'imaginer un modèle fixé a priori, c'est l'étude de la construction itérative (améliorations suc-

cessives du modèle et cercle vertueux pour la production de règles toujours plus intéressantes) qui est au centre de cet article.

## 2.3 Choix de l'utilisation d'un réseau bayésien pour l'analyse des règles

Notre hypothèse de départ est qu'il faut s'intéresser à la modélisation et l'exploitation des connaissances de l'expert pour faciliter l'analyse et l'interprétations des collections de règles d'association extraites. Cette direction de travail a donné lieu à plusieurs propositions. Padmanabhan et al. [15] ont étudié l'exploitation des connaissances de l'utilisateur explicitées par des règles. Cette approche a ensuite été formalisée par un réseau de croyances [16] permettant d'extraire l'ensemble minimum des règles d'association en fonction des connaissances du domaine. Ce type d'approche présente cependant une limite. En effet, une règle est jugée intéressante si elle diffère des règles définies dans le réseau de croyance, et non pas par rapport à ce que l'on pourrait inférer de ces croyances. Cette notion d'inférence a été développée dans [13]. Ces auteurs décrivent l'utilisation d'un réseau bayésien pour calculer l'intérêt d'ensembles d'attributs extraits à l'aide d'un algorithme de type Apriori [2]. La différence entre le support estimé sur les données et le support inféré à partir du réseau bayésien est calculée pour chaque ensemble d'attributs. Les motifs les plus intéressants sont ceux pour lesquels la divergence entre les connaissances de l'utilisateur (i.e., l'évaluation au moyen du réseau) et ce qui est observé dans les données réelles est la plus forte. Ces ensembles d'attributs sont ensuite soumis à l'utilisateur pour une éventuelle mise à jour de la structure ou des paramètres du réseau bayésien.

Nous choisissons également de modéliser les connaissances de l'expert à l'aide d'un réseau bayésien. Plus particulièrement, nous voulons exploiter certaines complémentarités intéressantes entre réseaux bayésiens et règles d'association :

- Liens de « dépendances » entre les variables (arcs du graphe, relations exprimées au sein des règles).
- Fréquence d'apparition des événements (tables de probabilités conditionnelles, support).

## 2.4 Exploitation du réseau bayésien

Soit $BD$ une base de données booléenne, et $H = \{A_1, A_2, \ldots, A_n\}$ l'ensemble de ses attributs booléens. $H$ est défini sur $D_H = D_{A_1} \times D_{A_2} \times \ldots \times D_{A_n}$. $P_I^{BD}(i)$ dénote la probabilité pour que l'ensemble d'attributs $I \subseteq H$ prenne comme valeur le vecteur $i$. Un itemset est représenté par la paire $(I, i)$ avec $I \in H$ ensemble d'attributs fini non vide et $i$ ensemble des valeurs des attributs de $I$. Lorsque cela n'est pas strictement nécessaire, l'itemset $(I, i)$ sera désigné simplement par $I$.

Un réseau bayésien $RB$ est un graphe dirigé acyclique défini par un ensemble de noeuds correspondants aux attributs de $H$ et par $E \subset H \times H$ l'ensemble des arcs du graphe. A chaque nœud on associe une distribution de probabilité conditionnelle $P_{A_i|\Pi_{A_i}}$, où $\Pi_{A_i} = \{A_j | (V_{A_j}, V_{A_i}) \in E\}$ représente les parents du nœud $A_i$. Pour une discussion détaillée sur les réseaux bayésiens le lecteur pourra consulter [18, 14]. Une des propriétés du réseau bayésien est de définir de manière unique la distribution de probabilité jointe de $H$ :

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

Une règle d'association $R$ est un motif $X \Rightarrow Y$, où $X$ et $Y$ sont des itemsets tels que $Y \neq \emptyset$ et $X \cap Y = \emptyset$. $X$ est appelé *partie gauche* de la règle, $Y$ la *partie droite*. Soit $I$ un itemset, le support de $I$ dans $BD$, noté $supp_{BD}(I)$, est l'ensemble des lignes (ou transactions) de $BD$ qui contiennent $I$.

Ainsi, étant donné une base de donnée $BD$ définie sur un ensemble d'attributs $H$ et un réseau bayésien $RB$, il est possible d'obtenir la confiance d'une règle d'association $R = X \Rightarrow Y$ (voir [6] pour un exemple d'algorithme d'inférence). En s'inspirant de [13], nous avons défini une mesure de l'intérêt d'une règle d'association. Cette mesure est basée sur la différence entre la confiance de la règle estimée à partir des données et celle inférée par le réseau bayésien. Pour une règle d'association $R = X \Rightarrow Y$ elle s'exprime de la manière suivante :

$$\begin{aligned} \mathrm{I}nt(R) &= |conf_{BD}(R) - conf_{RB}(R)| \quad (2) \\ \text{où } conf_{BD}(R) &= \frac{supp_{BD}(X \cup Y)}{supp_{BD}(X)} \\ \text{et } conf_{RB}(R) &= \prod_{i=1}^{m} P_{Y_i|\Pi_{Y_i}} \end{aligned}$$

Nous disposons donc 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.

## 2.5 Analyse des règles extraites

A l'issue de la phase de génération des règles d'association et du calcul de l'intérêt vis à vis du réseau bayésien, on peut utiliser la mesure d'intérêt (seuil $\epsilon$ défini par l'expert et souvent proche de zéro) pour classer les règles en deux groupes.

D'un côté, nous avons l'ensemble des règles d'association dont la valeur d'intérêt est inférieure au seuil $\epsilon$. Ces

règles n'apportent pas d'informations supplémentaires par rapport à ce qui est modélisé dans le réseau bayésien. L'analyste peut ignorer de telles règles, ce qui va faciliter et accélérer le processus d'analyse des résultats.

Le deuxième groupe de règles, dites $\epsilon$-intéressantes, contient les règles dont la valeur d'intérêt est supérieure au seuil $\epsilon$. Ces règles expriment des dépendances observées dans les données mais non explicitées dans le réseau, peuvent être de trois types :

- La règle contient une association connue de l'expert, mais non prise en compte par la modélisation du réseau bayésien. Il est alors important de modifier la structure du réseau bayésien 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.
- La règle est fortuite, elle représente 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 ».
- La règle est potentiellement intéressante. C'est à dire qu'elle « surprend » l'expert du domaine et va demander une analyse approfondie (e.g., nouvelle itération du processus de fouille, retour sur la collecte des données).

Nous voulons donc éliminer, après mise à jour des mesures d'intérêt, les motifs jusque là considérés comme $\epsilon$-intéressants mais qui sont déjà connus de l'expert. L'idée est d'affiner progressivement le modèle de connaissance utilisé pour le filtrage des règles d'association en y intégrant de nouvelles dépendances. Cependant, la modification et l'interprétation des motifs extraits ainsi que la modification appropriée du réseau bayésien ne sont pas faciles. Nous détaillons les difficultés rencontrées dans notre cas d'application à la section 3.4.

# 3 Application à des données d'interruptions opérationnelles

## 3.1 Présentation du cas d'application

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). Un tel événement est aujourd'hui considéré comme important par les compagnies aériennes du fait des coûts engendrés.

De ce fait, 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. Ces prédictions initient, guident et valident les choix de conception. Pour effectuer cette activité, les ingénieurs utilisent un outil informatique qui implémente un modèle mathématique stochastique intégrant les paramètres dont les impacts sur la fréquence des interruptions opérationnelles 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.

Les besoins de recherche portent sur l'amélioration des modèles de calcul utilisés par cet outil de prédiction. Dans ce contexte, la fouille des données en service est intéressante car elle permet de découvrir de nouveaux facteurs qui pourraient être intégrés à ces modèles pour améliorer la prédiction de la fréquence des interruptions opérationnelles. On se propose d'encadrer ce processus de découverte par l'approche méthodologique détaillée dans la section 2. L'analyse des données doit permettre de valider les hypothèses qui ont été prises et d'enrichir le modèle de prédiction.

Les sections suivantes présentent les différentes étapes du processus de fouille qui, appliqué selon la méthodologie proposée, permet de faciliter l'émergence de règles intéressantes, potentiellement exploitables -après reformulation- en tant que nouveaux contributeurs des taux d'interruptions opérationnelles. La base de données relative aux interruptions opérationnelles regroupe les détails de tous les problèmes techniques et elle a été définie en accord avec l'expert du domaine. Après pré-traitement, on dispose de 17 attributs discrétisés et de plus de 15000 enregistrements décrivant les interruptions opérationnelles. Un extrait de cette base de données est présenté dans le tableau 1.

## 3.2 Extraction des règles d'association

Dans un premier temps, nous avons regardé les résultats issus de l'extraction des règles d'association $\delta$-fortes. Ces règles sont présentées à l'expert en fonction de différentes mesures d'intérêt : confiance [1], J-mesure [19] et moindre contradiction [3]. Cette première extraction comporte de nombreuses règles connues de l'expert. Ceci permet de motiver le besoin en modèles de la connaissance a priori, qu'il s'agisse d'exploiter des dépendances exactes (taxonomies sur les composants) ou encore certaines croyances fortes de l'expert.

L'extraction a donné 17760 règles d'association. Le tableau 2 montre des exemples -choisis- de telles règles extraites au moyen de AC Miner [4] ($support_{min} = 100$, $\delta = 15$). Sur une configuration PC de bureau, l'extraction a demandé 2 minutes et 55 secondes.

Les mots-clés `remove`, `ecam`, `mel`, etc. indiquent que l'analyse du texte libre rédigé par un technicien a permis de déceler une action particulière : « pose/dépose » d'un équipement, apparition de messages d'alertes, application d'une procédure spécifique. Lorsqu'un mot-clé est préfixé de `last=` cela signifie qu'il s'agit du dernier mot-clé, correspondant à une action, détecté dans la description du problème. Les nombres de 2, 4 ou 6 chiffres désignent les équipements incriminés dans l'interruption opérationnelle.

| ATA Chapter | Date | Operator | MSN | Engine Type | Station | Phase | Effect code | Delay duration | Class | Text |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 29/12/1998 | OP1 | 11 | EngineXXA | ST3 | TX | DY/ / / | 0,5 | NM | LATE POSITIONNING EX MAINT. |
| 0 | 30/12/1998 | OP1 | 29 | EngineXXA | ST4 | CS | DY/ / / | 0,83 | NA | STR ASSY TO BE INSTALLED STR FIT |
| 212351 | 03/02/1998 | OP2 | 11 | EngineXXA | ST4 | CS | DY/ / / | 0,68 | | VENT EXTRACTION ECAM MESSAGE |
| 212600 | 07/10/1998 | OP1 | 50 | EngineXXA | ST1 | CS | DY/ / / | 0,39 | | AVIONIC VENT FAULT RETURN TO P |
| 212634 | 21/03/1998 | OP2 | 142 | EngineXXA | ST4 | TX | DY/ / / | 0,85 | | VENT OXXX VLV FAULT DURING TAXI |
| 212634 | 23/03/1998 | OP1 | 34 | EngineXXA | ST3 | CS | DY/ / / | 1,15 | | DURING PUSH BACK VENT EXTRACT |
| 212634 | 09/07/1998 | OP1 | 87 | EngineXXA | ST3 | CS | DY/ / / | 0,25 | | Vent extract fault. This caused a GXX |
| 212634 | 04/09/1998 | OP3 | 50 | EngineXXA | ST8 | TO | DY/ / / | 16 | NM | RTB DUE R CAB VENT FAULT MSG |
| 212634 | 13/09/1998 | OP4 | 42 | EngineXXA | ST2 | CS | DY/ / / | 2,37 | | SYS1 failed on puch back SYS1 reseted |
| 212651 | 07/09/1998 | OP3 | 151 | EngineXXA | ST1 | CS | DY/ / / | 0,51 | NS | AFTER ENG START ECAM WARNING |
| 212651 | 16/10/1998 | OP5 | 170 | EngineXXA | ST3 | CS | DY/ / / | 0,42 | | . |

**TAB. 1** – *Extrait de la base de données d'interruptions opérationnelles*

| Index | itemset $\delta$-libre $\Rightarrow$ fermeture | support | confiance |
|---|---|---|---|
| 1 | CS DY last=remove $\Rightarrow$ remove | 4046 | 1,00 |
| 2 | 2863 $\Rightarrow$ SYSTEM 28 286322 DY | 245 | 0,96 |
| 3 | TX last=nff $\Rightarrow$ SYSTEM DY nff | 247 | 0,94 |
| 4 | 15 $\Rightarrow$ ENGINE 1511 effect=DY | 178 | 0,96 |
| 5 | 028886 $\Rightarrow$ SYSTEM 02 0288 DY | 234 | 0,96 |
| 6 | ST2 last=none $\Rightarrow$ OP2 EngineXXA DY | 238 | 0,96 |
| 7 | 4345 delay=0.5_1.5 $\Rightarrow$ SYSTEM 43 434512 DY | 107 | 0,97 |
| 8 | OP1 EngineXXA CS delay=0.5_1.5 remove $\Rightarrow$ ST1 DY last=remove | 168 | 0,92 |
| 9 | 9911 TX delay=0.5_1.5 $\Rightarrow$ SYSTEM 99 991112 DY | 155 | 0,98 |
| 10 | 8885 month=Aug $\Rightarrow$ SYSTEM 88 DY | 183 | 0,95 |

**TAB. 2** – *Quelques règles d'association extraites*

Ces nombres obéissent à une taxonomie bien précise : la norme ATA 100. Ainsi l'équipement `286322` est un sous-équipement de `2863`, etc. Ces équipements se regroupent en trois catégories représentées par les mots-clés `SYSTEM`, `ENGINE` et `STRUCTURE`. Les codes `CS`, `TX`, indiquent la phase de vol pendant laquelle le problème est survenu (phase de vérification au sol, décollage, etc.). Les codes `DY` ou `CN` représentent la nature de l'interruption provoquée (retard, annulation, etc.). Pour des raisons de confidentialité, les données ont été falsifiées. Les sigles des compagnies et des aéroports, ainsi que les numéros de série des avions ont été volontairement rendus anonymes (`ST` fait référence aux aéroports et `OP` aux compagnies).

## 3.3 Construction du réseau bayésien initial

Dans notre contexte, le réseau bayésien représente les connaissances de l'expert qui vont permettre d'éliminer des motifs inintéressants. Dans un premier temps, nous allons modéliser les principales dépendances du domaine, i.e., des dépendances bien connues de l'expert. Pour ce faire, on passe généralement par trois étapes distinctes :

1. Identification des variables et de leur espace d'état.
2. Définition de la structure du réseau bayésien.
3. Définition de la loi de probabilité conjointe des variables.

**Identification des variables et de leur espace d'état.** Cette première étape exige une intervention « humaine », afin de déterminer l'ensemble des variables et leurs espaces d'état. Du point de vue de la fouille de données, c'est l'étape cruciale de la sélection et/ou de la construction de descripteurs.

Dans notre cas d'application, les variables retenues ont été déterminées à l'issue de plusieurs discussions avec l'expert. Elles déterminent le schéma de la base de données initiale pour le processus de fouille. La découverte de connaissances à partir de règles d'association suppose que les données capturent des relations booléennes. Les variables numériques doivent alors subir des discrétisations (liées à l'expertise dans, e.g., le cas de la variable `DELAY` qui représente le retard occasionné, ou dérivées tout simplement aux distributions rencontrées) pour finalement donner lieu à des ensembles de variables booléennes (une par intervalle de discrétisation retenu).

Notons que, au gré des itérations du processus de fouille, nous pouvons toujours enrichir la base de données en y ajoutant de nouvelles variables.

**Définition de la structure du réseau.** La deuxième étape consiste à identifier les liens entre les variables, i.e., dire pour quels couples de valeurs $(i, j)$ la variable $X_i$ influence-t-elle la variable $X_j$ ?

Le problème de l'apprentissage automatique de la structure est un problème difficile. Néanmoins de nombreux algorithmes ont été proposés (voir, e.g., [10] pour une présen-

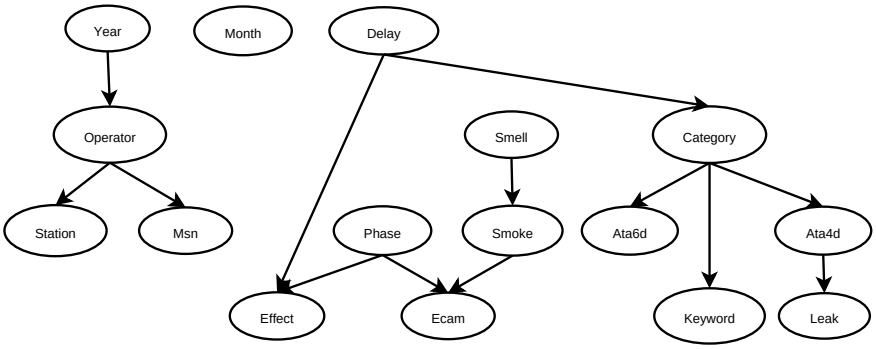


**FIG. 2** – *Réseau bayésien initial*

tation synthétique).

Si ces algorithmes fournissent une première structure de réseau bayésien, le résultat va cependant rarement convenir -tel quel- à l'expert du domaine. Ainsi, pour améliorer le processus d'apprentissage, il est courant de prendre en compte certaines connaissances de l'expert pour initier l'algorithme, puis pour modifier et valider le résultat obtenu.

Les différents *a* priori que peut vouloir exprimer l'expert sur les variables du réseau bayésien peuvent se résumer aux points suivants [5] :

- (1) Déclaration d'un nœud racine.
- (2) Déclaration d'un nœud feuille.
- (3) Déclaration d'un nœud cible (pour des applications de classification).
- (4) Déclaration d'un ordre (partiel ou complet) sur les variables.

La première structure ainsi obtenue est ensuite soumise au jugement de l'expert. Celui-ci peut alors décider de réviser plusieurs points :

- (5) Existence (ou absence) d'un arc entre deux nœuds .
- (6) Inversion du sens d'un arc entre deux nœuds précis.

Dans notre cas d'application, nous avons procédé à une recherche de structure à l'aide du logiciel Weka [20]. Cette première structure a été présentée à l'expert, ce qui lui permet dans un premier temps de visualiser les principales relations découvertes entre les variables. Il peut alors se concentrer sur la validation de cette structure, plutôt que de devoir l'élaborer en partant de rien. Cet aspect prend toute son importance dès que le nombre de variables à manipuler est élevé. Le réseau bayésien obtenu à l'issue de cette étape est présenté dans la figure 2.

**Définition des lois de probabilités conjointes des variables.** Cette étape consiste à définir les probabilités associées aux différents nœuds du graphe. L'apprentissage automatique des tables de probabilités est aujourd'hui bien maîtrisé [12], à condition que les données dont on dispose soient suffisamment représentatives du domaine. Malheureusement, dans de nombreuses applications réelles, il n'existe pas (ou très peu) de données disponibles. Il se peut aussi que les données soient en nombre suffisant mais qu'elles ne soient pas assez représentatives du domaine d'application. Dans ces situations, l'apprentissage des paramètres du réseau bayésien passe par l'utilisation des connaissances d'experts pour tenter d'estimer les probabilités conditionnelles.

Dans notre contexte, on dispose d'un seul expert, qui est jugé fiable, disponible et parfaitement familier avec les notions de probabilités. Le principal problème pour réaliser cette étape est le grand nombre de valeurs possibles à prendre en compte pour certaines variables. Ainsi, on se propose d'effectuer un apprentissage automatique des différentes tables de probabilités, à partir de l'ensemble des données disponibles, puis de soumettre cet apprentissage à l'expert. Celui-ci peut alors réviser certaines probabilités d'événements en fonction de ses connaissances. Pour cela on s'appuiera sur les travaux de [8] qui proposent la mise en place d'une *échelle de probabilité*. Celle-ci permet aux experts de faire la correspondance entre les probabilités « verbales » et numériques en assignant un degré de réalisation à une affirmation donnée, puis de comparer les probabilités des événements pour les modifier.

## 3.4 Modification du réseau bayésien

Nous avons donc un réseau bayésien qui a été initialement conçu par apprentissage automatique (figure 2). La structure ainsi que les tables de probabilités conjointes ont été validées par l'expert du domaine. On a aussi extrait une collection concise de règles d'association. Enfin, la mesure d'intérêt proposée permet de déterminer la divergence entre l'information représentée par une règle d'association donnée, et l'information que l'on peut inférer à partir du réseau bayésien.

Il est important de noter que le premier réseau bayésien construit n'a pas pour objectif d'être le plus complet possible par rapport aux connaissances de l'expert : il s'agit en fait de capturer les dépendances les plus évidentes, c'est-à-dire celles que l'on va retrouver le plus souvent dans les données. Puis, au fur et à mesure des itérations de notre pro-

| Index | Règle d'association | Annotation |
|---|---|---|
| 1 | SYSTEM ST2 last=remove ⇒ OP1 MB=true | (ST2 => OP1 ; C :0,8) (ST2 et OP1 => MB=true ; C :1,0) |
| 2 | 2851 ⇒ SYSTEM 285134 CS | (ATA4D => CATEGORY ; C :1,0) |
| 3 | msn=041 ⇒ SYSTEM OP3 | (MSN=041 => OP3 ; C :0,98) |
| 4 | OP3 MB=other ⇒ ST3 | (OP3 => ST3 ; C :0,65) (OP3 et ST3 => MB=other ; C :1,0) |
| 5 | month=1 OP4 ST1 ⇒ MB=true | (month => MB ; NV) (OP4 et ST1 => MB=true ; C :1,0) |
| 6 | CN leak last=remove ⇒ delay=6_inf | (CN => delay=6_inf ; C :1,0) (leak => delay=6_inf ; NV) |
| 7 | SYSTEM CS delay=6_inf ⇒ CN | (delay=6_inf => CN ; C :O,8) (CS => CN ; NV) |
| 8 | 361152 last=remove ⇒ SYSTEM 3611 | (361152 => SYSTEM ; C :1,0) (361152 => 3611 ; C :1,0) |
| 9 | ST2 last=remove ⇒ OP1 MB=true | (last=remove => OP1 ; NV) |
| 10 | 2793 ⇒ SYSTEM 279334 | (2793 => SYSTEM ; C :1,0) |

**TAB. 3** – *Exemple de règles d'association annotées*

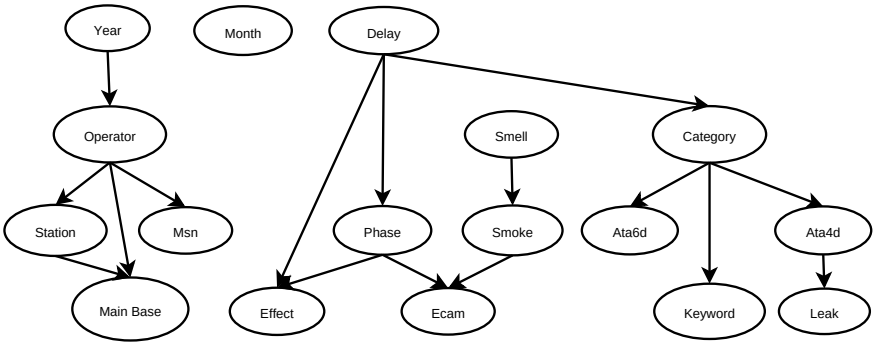


**FIG. 3** – *Réseau bayésien obtenu à l'issue du processus de fouille*

cessus de fouille, ce premier réseau sera modifié et complété pour capturer plus de connaissances du domaine et ainsi permettre l'émergence de règles d'association à la fois surprenantes et valides. Ainsi, à chaque itération du processus, les résultats de l'analyse des motifs par l'expert vont permettre de modifier le réseau afin d'intégrer différents types d'informations. Les motifs qui vont engendrer des modifications du réseau bayésien se divisent en trois catégories :

(C) Les règles d'associations font apparaître une dépendance connue, non intéressante et non prise en compte par le réseau bayésien.

(NV) Les règles d'associations font apparaître une dépendance non valide et non prise en compte par le réseau bayésien.

(I) Les règles d'associations sont jugées surprenantes et valides par l'expert. Elles représentent une connaissance utile qui sort du cadre de celles décrites par le réseau bayésien.

Dans des processus réels, le nombre des règles d'associations extraites qui appartiennent à ces catégories peut être très important. De plus, une association peut être composée à la fois d'une information connue et d'une information potentiellement intéressante, ce qui rend encore plus difficile -et donc sujette à erreurs- la tâche d'analyse des résultats de l'extraction par l'expert.

Afin de filtrer l'impact de ces deux types d'associations à la prochaine itération du processus de fouille, on demande à l'expert d'apporter un jugement sur les premiers motifs qu'il analyse en les annotant de manière bien précise. Ces annotations sont ensuite exploitées pour mettre à jour la structure et les paramètres du réseau.

Les catégories (NV) et (I) regroupent toutes les deux des motifs surprenants, la différence entre les deux passe par le jugement de l'expert et les investigations qu'il effectue sur les données pour confirmer ou non la validité du motif en question.

**Annotation des règles d'association.** Le principal problème par rapport à l'annotation des règles d'association est la définition d'une syntaxe qui soit à la fois suffisamment expressive pour l'expert et qui puisse se traduire le plus fidèlement possible dans la représentation du réseau bayésien. De fait, on a souhaité privilégier une syntaxe simple afin d'éviter toute confusion quant à l'interprétation et l'exploitation des annotations de l'expert. On peut résumer la syntaxe adoptée par la grammaire BNF suivante :

```
liste-annotation ::=
 liste-annotation annotation
 annotation

annotation ::=
```

```
 (prémisse => élément ; catégorie)

prémisse ::=
 liste-élément et élement
 élément

élément ::= one of
 attribut attribut=valeur

catégorie ::=
 C:probabilité-verbale
 NV
 I
```

Cette notation nous permet :

- De spécifier si une règle d'association contient aux yeux de l'expert un ou plusieurs motifs connus (C), jugés non valides (NV) ou le cas échéant, de marquer ce motif comme étant intéressant (I).
- De préciser clairement et sans ambiguïté la forme des motifs incriminés (liste de motifs simples dont la partie droite ne peut avoir qu'un seul élément).
- D'être générique quant à la forme des motifs détectés par l'expert (utilisation des noms d'attributs ou du couple attribut-valeur).
- De préciser, le cas échéant, la conjonction d'attributs ou d'items dans la prémisse du motif.
- D'associer une probabilité-verbale aux motifs définis comme connus. Pour cela on s'appuiera sur le principe de l'échelle de probabilité [7].

Le tableau 3 présente quelques exemples d'annotations effectuées sur des règles d'associations issues de notre cas d'application.

**Exploitation des annotations pour modifier le réseau.** *Considérons d'abord le cas des annotations de type (C).* 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 $P(Z|XY) = p$. On procédera de la même façon pour traiter les associations simples de type $X \Rightarrow Y$.

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 notre cas d'application, certaines variables (`OPERATOR`, `MSN`, `STATION`) ont une centaine de valeurs possibles ce qui rend toute modification manuelle de la structure du réseau relative à ces variables (par exemple la création d'un lien entre les variables `OPERATOR` et `MAIN BASE`) extrêmement coûteuse et délicate.
La solution appliquée 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) est actuellement 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 $A \Rightarrow B$.

Cette utilisation fait sens car l'expert a manifesté à plusieurs reprises le fait que les motifs non valides viennent gêner la lecture des résultats de l'extraction. Il faut cependant que l'expert soit excessivement prudent quant à l'utilisation de ce type de filtre pour ne pas cacher d'informations qui auraient pu se révéler intéressantes. Par la suite, il pourra être intéressant de réfléchir aux possibilités d'intégration de cette catégorie d'annotations par le biais de modifications appropriées du réseau bayésien.

Nos expérimentations préliminaires n'ont pour l'instant pas abouti à la découverte de nouvelle connaissance, c'est-à-dire à l'annotation de motifs comme étant à la fois surprenants et valides (I). Une réflexion est cependant nécessaire quant à la reformulation et à l'intégration de ce type de motifs dans notre modèle de connaissance. Comment doivent-ils être annotés par l'expert ? Comment les intégrer au réseau bayésien ? Comment gérer et garder une trace de ces nouvelles connaissances ? Ces questions restent pour l'instant des problèmes ouverts.

## 3.5 Résultats obtenus

La figure 4 présente une vue globale de la méthode proposée. On peut la résumer par les points suivants :

1. Apprentissage automatique d'un premier réseau bayésien, soumis à l'expert pour validation et modification.
2. Extraction, avec AC Miner, d'une collection concise de règles d'association à des seuils (fréquence, nombre d'exceptions) définis par l'utilisateur.
3. Calcul de l'intérêt de ces règles vis à vis du réseau bayésien (équation 2).

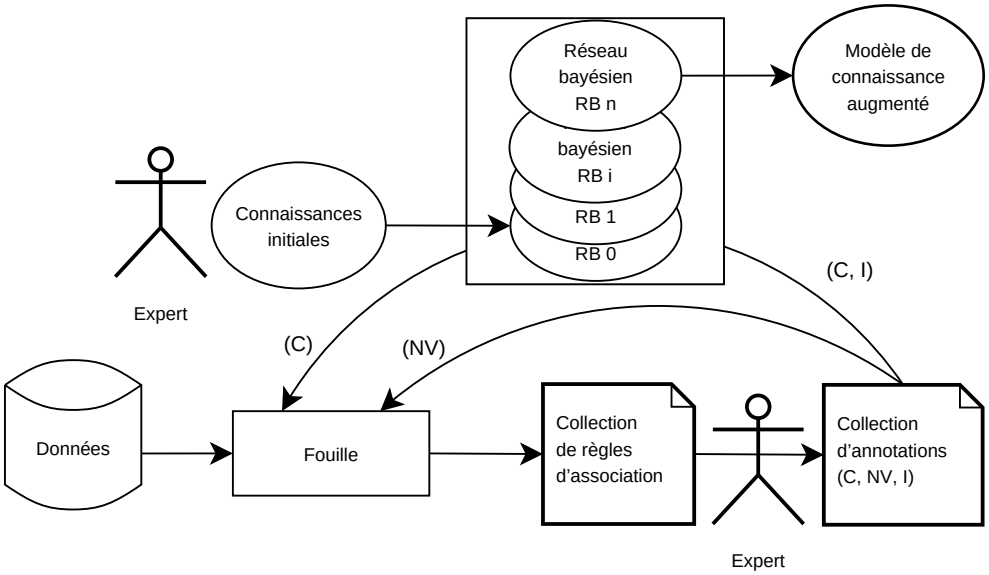


**FIG. 4** – *Exploitation des annotations pour la construction d'un réseau bayésien*

4. Classement des résultats obtenus par intérêt décroissant, et annotation des règles par l'expert.

5. Prise en compte des annotations par la mise à jour de la structure du réseau.

6. Nouvelle itération du processus à partir du point de 2 ou 3, selon que l'on a ajouté de nouvelles variables ou non.

Nous avons appliqué cette démarche expérimentale pour notre cas d'application. Le réseau bayésien obtenu au final est présenté dans la figure 3. En le comparant avec le réseau construit initialement (figure 2), on s'aperçoit que notre processus a permis de mettre en avant la construction de plusieurs arcs qui n'avaient pas été « trouvés » lors de la définition du premier réseau. Suite à l'analyse des premiers résultats de fouille, on a aussi pu introduire une nouvelle variable dans la base de données. Ainsi, les annotations rédigées pour les règles (1), (5) et (6) de notre exemple (figure 3) ont permis de détecter puis de modéliser l'influence conjointe des variables `OPERATOR` et `STATION` sur la variable `MAIN BASE`, en tant que motif connu de l'expert mais non initialement modélisé dans le réseau.

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.

Néanmoins, d'autres investigations sont nécessaires pour consolider ces premiers résultats et permettre la découverte de connaissances à partir de règles d'associations intéressantes.

# 4 Conclusion et travaux futurs

L'approche proposée a pour but de faciliter la tâche d'analyse des résultats d'extraction de règles d'association. Cette approche est basée sur l'extraction d'une collection de règles non redondantes aux propriétés intéressantes, l'utilisation de réseaux bayésiens pour la modélisation des dépendances connues du domaine d'application, l'utilisation d'une mesure d'intérêt qui prend en compte les connaissances de l'expert, et la mise en place d'un processus itératif qui permet de consolider et d'augmenter le modèle de connaissance initial mais aussi de faciliter l'émergence de motifs intéressants.

Les premiers résultats obtenus sur le cas d'application sont encourageants. Le prochain objectif fixé en termes d'expérimentations est de poursuivre les itérations de notre processus pour pouvoir le valoriser par la découverte de connaissances réellement surprenantes et utiles pour l'utilisateur.

Un axe de recherche 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) ?

## Références


[1] R. Agrawal, T. Imieliński et A. Swami. Mining association rules between sets of items in large databases. In P. Buneman et S. Jajodia, éditeurs, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington, D.C., 26–28 1993.

[2] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen et A. Inkeri Verkamo. *Fast discovery of association rules*, chapitre 12, pages 307–328. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996.

[3] J. Azé. *Extraction de connaissances à partir de données numériques et textuelles*. Thèse de doctorat, Université Paris-Sud, december 2003.

[4] J.-F. Boulicaut, A. Bykowski et C. 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.

[5] J. Cheng, R. Greiner, J. Kelly, D. Bell et W. Liu. Learning bayesian networks from data : An information-theory based approach. *Artificial Intelligence*, 137 :309–347, 2002.

[6] R. Dechter. Bucket elimination : A unifying framework for reasoning. *Artificial Intelligence*, 113(1-2) :41–85, 1999.

[7] M. J. Druzdzel et F. Diez. Criteria for combining knowledge from different sources in probabilistic networks, 2000.

[8] M. J. Druzdzel et L. 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.

[9] C. Fauré, S. Delprat, A. Mille et J.-F. 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.

[10] O. François et P. Leray. Etude comparative d'algorithmes d'apprentissage de structure dans les réseaux bayésiens. *Journal électronique d'intelligence artificielle*, 5(39) :1–19, 2004.

[11] B. Goethals et M. J. Zaki. *Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations FIMI 2003*. Melbourne, USA, 2003.

[12] D. Heckerman. A tutorial on learning with bayesian networks. Technical report, Microsoft Research, 1995.

[13] S. Jaroszewicz et D. 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.

[14] P. Naïm, P.-H. Wuilleminn, P. Leray, O. Pourret et A. Becker. *Réseaux bayésiens*. Eyrolles, 05 2004.

[15] B. Padmanabhan et A. 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.

[16] B. Padmanabhan et A. 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.

[17] N. 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.

[18] J. Pearl. *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[19] P. Smyth et R. M. Goodman. An information theoretic approach to rule induction from databases. *IEEE Transactions on Knowledge and Data Engineering*, 4(4) :301–316, 1992.

[20] I. H. Witten et E. Frank. *Data Mining : Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition edition, 2005.