Découverte et représentation des trajectoires de soins par analyse formelle de concepts

Nicolas Jay

Département de formation doctorale en informatique École doctorale IAEM Lorraine
UFR STMIA

Découverte et représentation des trajectoires de soins par analyse formelle de concepts

THÈSE

présentée et soutenue publiquement le 7 Octobre 2008

pour l'obtention du

Doctorat de l'université Henri Poincaré – Nancy 1
(spécialité informatique)

par

Nicolas Jay

Composition du jury

Président :	Marie-Christine Haton, Professeur, Université Henri Poincaré – Nancy 1
Rapporteurs :	Marius Fieschi, Professeur, Université de la Méditerranée
	Michel Habib, Professeur, Université Paris Diderot – Paris7
Examinateurs :	Anita Burgun, Professeur, Université de Rennes 1
	Bruno Crémilleux, Professeur, Université de Caen
Directeurs :	Amedeo Napoli, Directeur de recherche CNRS
	François Kohler, Professeur, Université Henri Poincaré – Nancy 1

Laboratoire Lorrain de Recherche en Informatique et ses Applications – UMR 7503

Remerciements

Mes remerciements s'adressent en premier lieu aux membres de mon jury : au professeur Marie-Christine Haton, qui me fait l'honneur de présider ce jury, ainsi qu'aux professeurs Anita Burgun et Bruno Crémilleux qui ont accepté de juger ce travail. Je remercie également les professeurs Marius Fieschi et Michel Habib, mes rapporteurs, pour leur lecture attentive de ce mémoire et leurs précieuses critiques.

Je dois beaucoup à mes deux directeurs de thèse, Amedeo Napoli et François Kohler. Ils témoignent d'un intérêt de longue date pour la problématique de cette thèse et se sont toujours montrés disponibles. J'ai eu la chance de bénéficier à la fois de la richesse complémentaire de leurs conseils et d'une grande liberté dans mes recherches. Au delà de leur apport scientifique, leurs qualités humaines et leur indéfectible soutien ont largement contribué à l'aboutissement de ce travail. Qu'ils reçoivent ici toute ma gratitude. J'espère que nos trajectoires suivront la même direction de nombreuses années encore ...

Un grand merci également à tous les membres l'équipe orpailleur dans laquelle il fait bon vivre et travailler. Merci à Jean pour ses conseils et ses bonnes blagues, à Amine toujours partant pour un café, à Yannick, Manu, Marie-Dominique, Malika, Nizar, Mehdi, Laszlo, Rokia, Adrien, Julien, Fadi, Florent, Mathieu, Sandy, Emmanuelle et tous les autres pour l'ambiance de travail et la bonne humeur.

Ma reconnaissance va à Madame le docteur Chantal Kohler, pour sa gentillesse et son soutien en de nombreuses occasions, et pour avoir pris le temps de relire ce document.

Toute mon amitié à l'équipe du laboratoire SPI-EAO pour ses encouragements, à Laurence, Alban, Anne, Luc, Hippolyte, Joelle et les anciens.

J'ai une pensée toute particulière pour mon ami Gilles, qui m'a un jour parlé de fouille de données ...

Merci à ceux de mes amis qui n'entendent rien en informatique et en santé publique, mais qui ont malgré tout essayé de comprendre ce en quoi mon travail consistait.

Je dédie ce travail à mes parents, mes frères et sœurs et mes grands parents pour leur affection et leur soutien.

Enfin, et surtout, je remercie Lucie, Jeanne et Célestine, pour leurs concessions et leur patience durant toutes ces années, pour leur amour et leur joie, et parce qu'elles sont le sens de ma vie.

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Trajectoires de soins

Parmi les grands enjeux du 21ème siècle, les systèmes de santé des pays occidentaux vont devoir faire face aux défis posés par l'augmentation de l'espérance de vie. Outre les questions relatives au vieillissement de la population, la prise en charge des maladies chroniques s'est considérablement modifiée au cours des cinquante dernières années et les progrès de la médecine ont permis d'améliorer le pronostic et la qualité de vie des patients atteints d'affections chroniques. En France, ce sont près de 8 millions de personnes qui sont prises en charge par l'assurance maladie au titre d'affection longue durée, c'est-à-dire des affections nécessitant un traitement prolongé et des soins particulièrement coûteux. On peut citer par exemple le diabète ou les cancers. Ces affections constituent un véritable enjeu de santé publique aux implications médicales, économiques, sociétales et humaines dans la mesure où il s'agit de maladies graves et où les patients touchés sont les plus fragiles. D'un point de vue institutionnel, les maladies chroniques nécessitent très souvent une approche multidisciplinaire. Un patient diabétique verra régulièrement au cours de son suivi non seulement son médecin traitant, mais aussi un endocrinologue, un biologiste, un ophtalmologiste, un podologue et, selon la gravité de sa maladie, un cardiologue, un neurologue... Les patients chroniques sont donc en contact avec de multiples acteurs du système de soins. Sur le plan médical, ils bénéficient d'une succession d'examens diagnostiques, de traitements médicaux ou chirurgicaux, et leur état de santé suit un cours variable : stabilité, aggravation, amélioration, crises, maladies intercurrentes. Il s'agit là d'une véritable *trajectoire de soins* dont les facettes sont multiples : géographique, temporelle, institutionnelle, médicale, économique ou sociale. La maîtrise des trajectoires de soins par le patient et ses soignants est importante à plus d'un titre : elle est un gage de qualité des soins, de qualité de vie pour le patient, et d'efficience médico-économique pour le système de soins. Elle repose sur la délivrance adaptée des thérapeutiques, l'observance du patient, un suivi

optimal, la non redondance des examens et la prévention des complications. Ces déterminants dépendent étroitement de la coordination des acteurs et de la planification des soins. L'approche multidisciplinaire amène les professionnels et les organismes de soins à travailler en réseau. Au mieux, les patients sont pris en charge au sein de filières de soins que l'on peut voir comme une prédétermination de leur trajectoire de soins. Pour faciliter la tâche des médecins, des groupes d'experts élaborent des guides de bonnes pratiques cliniques qui sont des recommandations actualisées issues de la recherche médicale. Leur objectif est de standardiser les prises en charge à partir d'une médecine fondée sur des faits scientifiquement prouvés. Ces guides peuvent se traduire en programmes de soins afin d'orienter au mieux les trajectoires individuelles des patients. Néanmoins, il n'est pas évident que ces recommandations soient suivies en pratique. De surcroît, dans une prise en charge multidisciplinaire, des facteurs organisationnels peuvent être à l'origine de dysfonctionnements dans la mise en œuvre des programmes de soins.

Dans une optique d'évaluation des soins, il est indispensable d'observer, comprendre et modéliser les trajectoires de soins dans leur réalité. Pour mener à bien ce processus, il faut disposer d'informations sur le parcours des patients dans le système de soins. À l'échelle individuelle, cette information est morcelée et répartie dans des sources multiples : dossiers papier ou informatisés du généraliste, des spécialistes, dossier papier hospitalier par service, système d'information hospitalier ... L'informatisation des données médicales, l'interopérabilité des systèmes d'information, le déploiement du futur dossier médical personnel devraient résoudre peu à peu cette question. Cependant, si les données existent, très peu d'outils à ce jour sont capables de les mobiliser pour décrire et classifier un ensemble de trajectoires de soins. L'objectif principal de nos travaux est donc de concevoir une méthode d'analyse des trajectoires de soins dans une population à partir de données de systèmes d'information existants et qui n'ont pas été nécessairement conçus pour cette tâche.

Extraction de connaissances à partir de données

Depuis vingt ans, les progrès effectués dans la collecte et le stockage des données ont conduit à la création de nombreuses et volumineuses bases de données. Le rôle premier de ces bases de données est d'assurer la pérennité et la cohérence des données. Ces dernières restent essentiellement utilisées en consultation ou pour générer des rapports descriptifs simples. Dans certains domaines, comme celui de la recherche médicale, l'analyse des données fait appel à des méthodes statistiques plus élaborées. Il demeure dans ce cas que le format et l'organisation des données recueillies sont adaptés à une stratégie d'analyse définie au préalable. Il apparaît néanmoins que des éléments potentiellement utiles peuvent se nicher au cœur des bases de données. Ces éléments peuvent être des liens particuliers entre les données qui présentent un

intérêt pour les experts du domaine relatif à ces données. Par exemple, de nombreux hôpitaux disposent de systèmes de prescription informatisée. Par ailleurs, ils enregistrent de manière systématique des données relatives à l'état de santé des patients. Ces deux bases de données font fréquemment l'objet d'exploitations statistiques séparées dans un but de gestion. Or, elles contiennent ensemble suffisamment d'informations pour rechercher des incidents thérapeutiques qui ne pourront être mis en évidence que par un processus actif de recherche.

Une base de données peut alors être vue comme une mine d'information brute qu'il faut creuser et fouiller pour en extraire des pépites de connaissances. Pour autant, la complexité des données et leur grand volume rendent cette extraction non triviale. Il est inenvisageable d'effectuer cette tâche de manière complètement manuelle et empirique. C'est la raison pour laquelle ont été développés des méthodes et des outils dédiés capables d'assister un analyste dans sa découverte de connaissances. La conception de ces méthodes a donné naissance à un courant de recherche multidisciplinaire qui rassemble des experts des bases de données, des statistiques, de l'apprentissage, du raisonnement et de la représentation des connaissances : l'*Extraction de Connaissances à partir de Données* (ECD), ou en anglais *Knowledge Discovery in Databases*. La méthode que nous avons mise au point pour l'analyse des trajectoires de soins s'inscrit au cœur de l'ECD et repose sur une méthode de classification : l'Analyse Formelle de Concepts.

Analyse Formelle de Concepts

L'Analyse Formelle de Concepts (AFC) est une méthode d'analyse de données et de représentation des connaissances qui traite l'information sous la forme de hiérarchies de concepts construites à partir de tableaux binaires. L'AFC est une méthode de classification non supervisée. Elle est capable de découvrir et organiser les regroupements naturels d'un ensemble d'objets lié à un ensemble d'attributs. Elle est particulièrement adaptée au traitement de données symboliques. Un de ses principaux atouts réside dans ses propriétés de visualisation. L'AFC produit des treillis de concepts formels qui peuvent être représentés par un graphe et faciliter ainsi la tâche de l'analyste dans la découverte de connaissances. Notre méthode d'analyse des trajectoires de soins tire grandement parti de cette propriété.

L'AFC entretient des liens étroits avec d'autres méthodes d'ECD et en particulier la recherche de motifs fréquents. À partir d'un ensemble d'objets `O` caractérisés par un ensemble d'attributs (ou items) `I`, la recherche de motifs fréquents consiste à retrouver les sous-ensembles de `I` dont les éléments sont des attributs qui décrivent conjointement une proportion importante des objets de `O`. Les motifs fréquents sont à la base de la découverte de règles d'associations, objet de nombreux travaux de recherche. Notre intérêt pour les motifs fréquents découle de

leur équivalent en AFC : les concepts fréquents. Les treillis de concepts fréquents, ou icebergs, permettent de réduire la complexité des résultats de l'AFC en présence de grands volumes de données. Toutefois, fréquence n'est pas synonyme de pertinence. L'objectif de l'ECD est de mettre en évidence des connaissances qui présentent un intérêt pour l'analyste. La notion de fréquence s'appuie sur une mesure d'intérêt des motifs, ou des concepts : le support. Cette mesure se révèle parfois inadaptée au but poursuivi par l'analyste. C'est la raison pour laquelle nos recherches ont également porté sur une autre mesure d'intérêt des concepts formels : la stabilité. Cette mesure offre un point de vue différent sur les données et possède l'avantage d'être résistante au bruit. Dans ce travail, nous présentons une analyse qualitative de la stabilité et montrons son utilité pour la découverte de classes pertinentes de trajectoires de soins.

Enfin, la recherche de motifs fréquents a donné naissance à une extension pour l'analyse de données ordonnées : les motifs fréquents séquentiels. L'objectif ici est de découvrir des attributs qui apparaissent fréquemment ensemble et dans le même ordre chez les objets de la base de données. Typiquement, cette méthode est utilisée pour étudier des séquences d'évènements dans le temps. Elle s'avère donc intéressante pour l'analyse des trajectoires de soins. Nous présentons ici une approche originale de classification de motifs séquentiels par AFC. De plus, nous étudions l'effet de la simplification de treillis de motifs séquentiels par la mesure de stabilité.

Plan de la thèse

Le premier chapitre pose les fondements théoriques de l'AFC et décrit les principes de la visualisation de treillis de concepts. Il présente quelques extensions théoriques récentes de l'AFC puis aborde les principaux algorithmes de construction et manipulation de treillis, ainsi que les logiciels qui les implémentent. Dans la dernière section, nous dressons l'état des lieux des domaines d'application de l'AFC.

Dans le deuxième chapitre, nous présentons la recherche de motifs fréquents simples et séquentiels. En décrivant ses liens avec l'AFC, nous détaillons en particulier les icebergs de concepts.

Le troisième chapitre est consacré à l'étude des mesures d'intérêt de concepts formels. Nous y définissons la stabilité et en présentons une analyse qualitative détaillée.

Le quatrième chapitre situe l'analyse des trajectoires de soins dans le contexte médical. Après une présentation des processus d'acquisition et de mise en œuvre du savoir en médecine, nous définissons la notion de trajectoire de soins et rappelons les principaux systèmes de classification de malades existants. Le chapitre s'achève par une présentation du Programme de Médicalisation des Systèmes d'Information, qui est la source de données sur laquelle nous avons expérimenté notre méthode.

Dans le dernier chapitre, nous conduisons plusieurs expérimentations sur des données réelles pour illustrer l'intérêt de l'AFC dans l'étude des trajectoires de soins. Celles-ci sont précédées d'une introduction sur l'analyse de réseaux sociaux par AFC. La suite montre la richesse de la visualisation par treillis de concepts de l'organisation du système de soins en Lorraine, dans le domaine de la cancérologie. Nous mettons par ailleurs en évidence l'utilité des mesures de support et de stabilité des concepts pour réduire la complexité d'un treillis volumineux. Une partie du chapitre s'attache à montrer la capacité de l'AFC à produire des profils de morbidité ayant une cohérence clinique. Enfin, nous proposons une approche originale mettant en œuvre des treillis de motifs séquentiels fréquents pour établir une typologie des séquences de soins.

Nous terminons ce mémoire par un résumé des apports et une mise en perspective de nos travaux.

Chapitre 1

Analyse Formelle de Concepts

1.1 Introduction

L'Analyse Formelle de Concepts (AFC) est une méthode d'analyse de données et de représentation des connaissances qui traite l'information sous la forme de hiérarchies de concepts. L'AFC est née au début des années 1980. Bien que ses prémisses remontent aux années 1940 avec les travaux de Birkhoff [Birkhoff, 1940] sur les treillis, puis ceux de Barbut et Monjardet sur les treillis de Galois [Barbut and Monjardet, 1970] dans les années 1960, elle a été introduite sous ce nom par Wille [Wille, 1982]. Conçue à l'origine comme une restructuration de l'algèbre générale et de la théorie des treillis, elle s'est d'abord développée autour d'une communauté de mathématiciens. À partir des années 1990, elle a progressivement attiré les chercheurs en informatique et servi de fondement théorique à de nombreuses applications. Aujourd'hui, elle couvre des champs aussi divers que la linguistique, la recherche d'information, la fouille de textes, le génie logiciel, la représentation des connaissances et plus récemment l'extraction des connaissances à partir de données. Ce chapitre fait l'état de l'art de l'AFC. Il débute par une section consacrée aux fondements théoriques de l'AFC. La section suivante décrit les méthodes de visualisation des treillis de concepts. La troisième section s'intéresse aux extensions de l'AFC que sont les contextes multivalués, les treillis entrelacés et l'analyse relationnelle de concepts. Les deux sections suivantes abordent l'aspect informatique de l'AFC avec une présentation des principaux algorithmes et logiciels de construction et de manipulation des treillis de concepts. Le chapitre se termine par un état des lieux des applications les plus courantes de l'AFC.

1.2 Fondements théoriques

Les bases mathématiques de l'AFC ont été posées dans un livre de Ganter et Wille [Ganter and Wille, 1999]. Par la suite, nous nous conformerons aux notations utilisées dans cet ouvrage. Cependant, plusieurs formalismes coexistent et l'on parle aussi de Treillis de Galois.

1.2.1 Contexte formel

L'analyse formelle de concepts étudie l'organisation naturelle d'un ensemble d'objets en relation avec un ensemble d'attributs. Mathématiquement, cette relation peut être représentée par un contexte formel.

Définition 1 (Contexte Formel)**.** *Un contexte formel* $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$ *est en triplet où* $\mathtt{G}$ *est un ensemble d'objets formels,* $\mathtt{M}$ *un ensemble d'attributs formels et* $\mathtt{I}$ *une relation binaire dans* $\mathtt{G} \times \mathtt{M}$*. Pour tout couple* $(\mathtt{g}, \mathtt{m}) \in \mathtt{I}$ *on dit que l'objet* $\mathtt{g}$ *« possède » l'attribut* $\mathtt{m}$*.*

Un contexte formel peut être visualisé dans un tableau binaire. Le tableau 1.1 présente le contexte formel `pathologie` caractérisant un ensemble de patients et leurs pathologies. Les patients sont des objets formels et les pathologies des attributs formels. La présence d'une pathologie chez un malade est matérialisée par une croix à l'intersection de leur colonne et ligne respectives. Ainsi, le patient `p1` est atteint de `diabete` et d'`hypertension`.

	diabete	hypertension	dyslipidemie	angor
p1	×	×		
p2		×		
p3		×		
p4	×			
p5	×		×	×
p6	×			
p7	×	×	×	×
p8	×	×		

TAB. 1.1 – Le contexte formel `pathologie`.

Alors que $\mathtt{G}$ et $\mathtt{M}$ désignent généralement un ensemble d'objets, respectivement d'attributs, il est possible de permuter leurs rôles : si $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$ est un contexte formel, alors $(\mathtt{M}, \mathtt{G}, \mathtt{I}^{-1})$, avec $(\mathtt{g}, \mathtt{m}) \in \mathtt{I} \Leftrightarrow (\mathtt{m}, \mathtt{g}) \in \mathtt{I}^{-1}$ est son contexte dual.

1.2.2 Concepts formels

Dans un contexte formel, les concepts formels expriment les regroupements naturels des objets et de leurs propriétés. Pour définir la notion de concept formel, on introduit les opérateurs de dérivation suivants, tous deux notés $'$ (prime) :

Définition 2 (Opérateur de Dérivation). *Soit* $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$ *un contexte formel, on définit sur* $2^{\mathtt{G}}$ *l'ensemble des parties de* $\mathtt{G}$*, respectivement sur* $2^{\mathtt{M}}$ *l'ensemble des parties de* $\mathtt{M}$*, les fonctions :*

$$' : 2^{\mathtt{G}} \rightarrow 2^{\mathtt{M}} : \mathtt{X}' = \{\mathtt{m} \in \mathtt{M} | \forall \mathtt{g} \in \mathtt{X}, (\mathtt{g}, \mathtt{m}) \in \mathtt{I}\}$$

$$' : 2^{\mathtt{M}} \rightarrow 2^{\mathtt{G}} : \mathtt{Y}' = \{\mathtt{g} \in \mathtt{G} | \forall \mathtt{m} \in \mathtt{Y}, (\mathtt{g}, \mathtt{m}) \in \mathtt{I}\}$$

On pose également : $(\mathtt{X} = \emptyset)' = \mathtt{M}$ *et* $(\mathtt{Y} = \emptyset)' = \mathtt{G}$

En reprenant le contexte formel `pathologie` du tableau 1.1, on a :
- $\{\mathtt{p1}, \mathtt{p3}\}' = \{\mathtt{hypertension}\}$,
- $\{\mathtt{diabete}, \mathtt{dyslipidemie}\}' = \{\mathtt{p5}, \mathtt{p7}\}$

$\{\mathtt{hypertension}\}$ représente l'ensemble des pathologies qui sont communes aux patients `p1` et `p3`. De même, $\{\mathtt{p5}, \mathtt{p7}\}$ représente l'ensemble des patients qui sont atteints des pathologies $\{\mathtt{diabete}, \mathtt{dyslipidemie}\}$.

Définition 3 (Opérateur de Composition). *Soit* $\mathbb{K} = (G, M, I)$ *un contexte formel, on définit les opérateurs de composition et on note* $''$ *les fonctions suivantes :*

$$'' : 2^{\mathtt{G}} \rightarrow 2^{\mathtt{G}} : \mathtt{X}'' = (\mathtt{X}')'$$

$$'' : 2^{\mathtt{M}} \rightarrow 2^{\mathtt{M}} : \mathtt{Y}'' = (\mathtt{Y}')'$$

Dans le contexte formel `pathologie`, on a :
- $\{\mathtt{p1}, \mathtt{p3}\}'' = \{\mathtt{hypertension}\}' = \{\mathtt{p1}, \mathtt{p2}, \mathtt{p3}, \mathtt{p7}, \mathtt{p8}\}$,
- $\{\mathtt{angor}\}'' = \{\mathtt{p4}\}' = \{\mathtt{diabete}, \mathtt{angor}\}$

Définition 4 (Concept Formel). *On appelle concept formel d'un contexte formel* $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$ *un couple* $\mathtt{C} = (\mathtt{X}, \mathtt{Y}) \in \mathtt{G} \times \mathtt{M}$ *tel que :*

$$\mathtt{X}' = \mathtt{Y} \text{ et } \mathtt{Y}' = \mathtt{X}$$

$\mathtt{X}$ *est appelé extension du concept* $\mathtt{C}$*.* $\mathtt{Y}$ *est appelé intension du concept* $\mathtt{C}$*.*
On note :
- $\mathfrak{B}(\mathtt{G}, \mathtt{M}, \mathtt{I})$ *ou* $\mathfrak{B}(\mathbb{K})$ *l'ensemble des concepts formels du contexte* $\mathbb{K}$*,*

- $\mathfrak{U}(\mathbb{K})$ *l'ensemble des extensions de ces concepts,*
- $\mathfrak{J}(\mathbb{K})$ *l'ensemble des intensions de ces concepts.*

Notons que si $(\mathtt{X}, \mathtt{Y})$ est un concept formel, par définition on a $\mathtt{X}'' = \mathtt{X}$ et $\mathtt{Y}'' = \mathtt{Y}$.

Le couple $(\{\mathtt{p1}, \mathtt{p7}, \mathtt{p8}\}, \{\mathtt{diabete}, \mathtt{hypertension}\})$ est un concept formel du contexte `pathologie` (tableau 1.1). L'ensemble des maladies communes à `p1`, `p7` et `p8` est $\{\mathtt{diabete}, \mathtt{hypertension}\}$. L'ensemble des patients qui sont atteints à la fois de `diabete` et d'`hypertension` est bien $\{\mathtt{p1}, \mathtt{p7}, \mathtt{p8}\}$.

En revanche,$(\{\mathtt{p1}, \mathtt{p8}\}, \{\mathtt{hypertension}\})$ n'est pas un concept formel. En effet :

- $\{\mathtt{p1}, \mathtt{p8}\}' = \{\mathtt{hypertension}, \mathtt{diabete}\}$
- et $\{\mathtt{hypertension}\}' = \{\mathtt{p1}, \mathtt{p2}, \mathtt{p3}, \mathtt{p7}, \mathtt{p8}\}$.

Bien qu'un concept formel soit représenté par deux dimensions, son extension et son intension, celles-ci sont étroitement liées et chacune détermine l'autre ($\mathtt{X}' = \mathtt{Y}, \mathtt{Y}' = \mathtt{X}$). La définition d'un concept formel répond à un principe de maximalité et il n'est pas possible d'étendre l'extension d'un concept sans modifier son intension. De même il n'est pas possible d'étendre l'intension d'un concept sans modifier son extension. Cela peut être visualisé dans un contexte formel en permutant les lignes et les colonnes du tableau. Un concept formel apparaît alors comme un rectangle maximal rempli de croix. Le tableau 1.2 montre le concept $(\{\mathtt{p1}, \mathtt{p7}, \mathtt{p8}\}, \{\mathtt{diabete}, \mathtt{hypertension}\})$ du contexte `pathologie` sous la forme d'un rectangle maximal.

	diabete	hypertension	dyslipidemie	angor
p1	×	×		
p7	×	×	×	×
p8	×	×		
p2		×		
p3		×		
p4	×			
p5	×		×	×
p6	×			

TAB. 1.2 – Le concept $(\{\mathtt{p1}, \mathtt{p7}, \mathtt{p8}\}, \{\mathtt{diabete}, \mathtt{hypertension}\})$ est un rectangle maximal.

Définition 5 (Classe d'équivalence)**.** *Soit* $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$ *un contexte formel et* $\mathtt{X} \subseteq \mathtt{G}$*. On appelle*

classe d'équivalence de $\mathtt{X}$ *et on note* $\langle X \rangle$ *l'ensemble :*

$$\langle \mathtt{X} \rangle = \{\mathtt{Y} \subseteq \mathtt{G} | \mathtt{Y}' = \mathtt{X}'\}$$

On définit de même la notion de classe d'équivalence pour un ensemble d'attributs formels en remplaçant $\mathtt{G}$ *par* $\mathtt{M}$*.*

Reformulé en termes de classe d'équivalence, le principe de maximalité s'énonce ainsi : si $(\mathtt{X}, \mathtt{Y})$ est un concept formel, alors $\mathtt{X}$ est le maximum, au sens de l'inclusion, de sa propre classe d'équivalence $\langle \mathtt{X} \rangle$. De même, $\mathtt{Y}$ est le maximum de sa classe d'équivalence $\langle \mathtt{Y} \rangle$. L'ensemble des concepts d'un contexte formel $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$ peut être également vu comme une bijection entre une partition de $\mathtt{G}$ et une partition de $\mathtt{M}$. En effet :

Propriété 1. *Soient un contexte formel* $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$ *et* $\mathfrak{U}(\mathbb{K})$ *(respectivement* $\mathfrak{J}(\mathbb{K})$*) l'ensemble des extensions (respectivement intensions) de ses concepts formels. L'ensemble des classes d'équivalences générées par les éléments de* $\mathfrak{U}(\mathbb{K})$ *(respectivement intensions de* $\mathfrak{J}(\mathbb{K})$*) forme une partition de* $\mathtt{G}$ *(respectivement* $\mathtt{M}$*).*

Par ailleurs, chaque classe d'équivalence (d'objets ou d'attributs) peut être décrite par un ensemble d'éléments minimaux incomparables entre eux : les générateurs minimaux.

Définition 6 (Générateur minimal)**.** *Soit* $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$ *un contexte formel et* $\mathtt{X} \subseteq \mathtt{G}$ *(respectivement* $\mathtt{M}$*. Soit* $\theta \subseteq \mathtt{X}$*.* θ *est un générateur minimal de* $\langle \mathtt{X} \rangle$ *si et seulement si :*

$$\theta' = \mathtt{X}' \text{ et } \nexists \mathtt{Y} \subseteq \theta | \mathtt{Y}' = \mathtt{X}'$$

.

Il est possible de définir une relation d'ordre sur l'ensemble des concepts d'un contexte formel.

Définition 7 (Sous-concept, Super-concept)**.** *Soit* $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$ *un contexte formel et* $\mathfrak{B}(\mathbb{K})$ *l'ensemble de ses concepts formels. Soient* $\mathtt{C_1} = (\mathtt{X_1}, \mathtt{Y_1})$ *et* $\mathtt{C_2} = (\mathtt{X_2}, \mathtt{Y_2})$ *deux concepts de* $\mathfrak{B}(\mathbb{K})$*. On dit que* $\mathtt{C_1}$ *est un sous-concept de* $\mathtt{C_2}$ *et on note* $\mathtt{C_1} \leq \mathtt{C_2}$ *si et seulement si :*

$$\mathtt{X_1} \subseteq \mathtt{X_2}$$

Dans ce cas, $\mathtt{C_2}$ *est un super-concept de* $\mathtt{C_1}$*.*

Cette relation d'ordre sur les concepts peut être définie de manière équivalente à partir de l'inclusion des intensions.

Propriété 2. *Soient* $\mathtt{C}_1 = (\mathtt{X}_1, \mathtt{Y}_1)$ *et* $\mathtt{C}_2 = (\mathtt{X}_2, \mathtt{Y}_2)$ *deux concepts de* $\mathfrak{B}(\mathbb{K})$.

$$\mathtt{C}_1 \leq \mathtt{C}_2 \Leftrightarrow \mathtt{X}_1 \subseteq \mathtt{X}_2 \Leftrightarrow \mathtt{Y}_2 \subseteq \mathtt{Y}_1$$

1.2.3 Treillis de concepts

L'ensemble des concepts d'un contexte formel, muni de la relation d'ordre sur les concepts, possède une structure mathématique particulière de *treillis*. Avant de définir la notion de treillis de concepts, nous rappelons quelques définitions relatives aux ensembles ordonnés.

Définition 8 (Majorant, Minorant)**.** *Soit* $(\mathtt{E}, \leq)$ *un ensemble ordonné et* $\mathtt{S}$ *une partie de* $\mathtt{E}$. *On appelle majorant (respectivement minorant) de* $\mathtt{S}$ *tout élément* $\mathtt{a} \in \mathtt{E}$ *tel que :*

$$\forall \mathtt{x} \in \mathtt{S}, \mathtt{x} \leq \mathtt{a} \qquad \text{(respectivement } \mathtt{x} \geq \mathtt{a})$$

Définition 9 (Supremum,Infimum)**.** *Soit* $(\mathtt{E}, \leq)$ *un ensemble ordonné et* $\mathtt{S}$ *une partie de* $\mathtt{E}$. *On appelle supremum ou borne supérieure (respectivement infimum ou borne inférieure) de* $\mathtt{S}$*, s'il existe, et on note* $\bigvee \mathtt{S}$ *(respectivement* $\bigwedge \mathtt{S}$*) le plus petit des majorants (respectivement le plus grand des minorants) de* $\mathtt{S}$.

Pour une paire $(\mathtt{x}, \mathtt{y})$ d'éléments de $\mathtt{E}$, le majorant $\bigvee\{\mathtt{x}, \mathtt{y}\}$ (respectivement le minorant $\bigwedge\{\mathtt{x}, \mathtt{y}\}$) est noté $\mathtt{x} \vee \mathtt{y}$ (respectivement $\mathtt{x} \wedge \mathtt{y}$).

Définition 10 (Relation inverse)**.** *Soit* $(\mathtt{E}, \leq)$ *un ensemble ordonné. On appelle relation inverse de* $\leq$ *et on note* $\geq$ *la relation définie par :*

$$\forall (\mathtt{x}, \mathtt{y}) \in \mathtt{E}^2, \mathtt{x} \leq \mathtt{y} \Leftrightarrow \mathtt{y} \geq \mathtt{x}$$

Propriété 3 (Principe de dualité)**.** *Soit* $\geq$ *la relation inverse de* $\leq$. *Toute assertion sur* $\leq$,$\wedge$,$\vee$ *reste vraie en remplaçant* $\leq$ *par* $\geq$ *et en permutant* $\wedge$ *et* $\vee$.

Si $(\mathtt{E}, \leq)$ est un ensemble ordonné, alors $(\mathtt{E}, \geq)$ est également un ensemble ordonné. Le principe de dualité permet de déduire les propriétés duales de $(\mathtt{E}, \geq)$ à partir des propriétés de $(\mathtt{E}, \leq)$.

Définition 11 (Successeur direct, Prédécesseur direct)**.** *Soit* $(\mathtt{E}, \leq)$ *un ensemble ordonné et* $\mathtt{a}$ *et* $\mathtt{b}$ *deux éléments de* $\mathtt{E}$. $\mathtt{a}$ *est un successeur direct de* $\mathtt{b}$ *et on note* $\mathtt{a} \prec \mathtt{b}$ *si* $\mathtt{a} < \mathtt{b}$ *et s'il n'existe pas d'élément* $c \in \mathtt{E}$ *tel que* $\mathtt{a} < \mathtt{c} < \mathtt{b}$. *Dans ce cas,* $\mathtt{b}$ *est le prédécesseur direct de* $\mathtt{a}$.

Définition 12 (Treillis)**.** *Un ensemble ordonné* $(\mathtt{E}, \leq)$ *est un treillis si pour tout* $\mathtt{x}$ *et* $\mathtt{y}$ *de* $\mathtt{E}$, $\mathtt{x} \vee \mathtt{y}$ *et* $\mathtt{x} \wedge \mathtt{y}$ *existent.*

Définition 13 (Treillis Complet). *Un ensemble ordonné* $(\mathtt{E}, \leq)$ *est un treillis complet si pour toute partie* $\mathtt{S}$ *de* $\mathtt{E}$, $\bigvee \mathtt{S}$ *et* $\bigwedge \mathtt{S}$ *existent.*

Propriété 4 (Propriété fondamentale de l'AFC [Wille, 1982]). *Soit* $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$ *un contexte formel. On note* $\underline{\mathfrak{B}}(\mathtt{G}, \mathtt{M}, \mathtt{I})$ *ou* $\underline{\mathfrak{B}}(\mathbb{K})$ *l'ensemble* $\mathfrak{B}(\mathtt{G}, \mathtt{M}, \mathtt{I})$ *muni de la relation d'ordre* $\leq$. *Soit* $\mathtt{T}$ *un index de* $\underline{\mathfrak{B}}(\mathbb{K})$. $\underline{\mathfrak{B}}(\mathbb{K})$ *est un treillis complet appelé treillis de concepts de* $\mathbb{K}$ *dans lequel les infima et suprema sont donnés par :*

$$\begin{aligned} \bigwedge_{\mathtt{t}\in\mathtt{T}} (\mathtt{X}_\mathtt{t}, \mathtt{Y}_\mathtt{t}) &= (\bigcap_{\mathtt{t}\in\mathtt{T}} \mathtt{X}_\mathtt{t}, (\bigcup_{\mathtt{t}\in\mathtt{T}} \mathtt{Y}_\mathtt{t})'') \\ \bigvee_{\mathtt{t}\in\mathtt{T}} (\mathtt{X}_\mathtt{t}, \mathtt{Y}_\mathtt{t}) &= ((\bigcup_{\mathtt{t}\in\mathtt{T}} \mathtt{X}_\mathtt{t})'', \bigcap_{\mathtt{t}\in\mathtt{T}} \mathtt{Y}_\mathtt{t}) \end{aligned}$$

Le supremum de tous les concepts de $\underline{\mathfrak{B}}(\mathbb{K})$ *est également noté* $\top$ *(top), leur infimum* $\bot$ *(bottom).*

Définition 14 (Couverture supérieure, Couverture inférieure). *La couverture supérieure (respectivement inférieure) d'un concept est l'ensemble de tous ses prédécesseurs directs (respectivement successeurs directs).*

Définition 15 (Filtre, Idéal). *Soit* $\underline{\mathfrak{B}}(\mathbb{K})$ *un treillis de concepts et* $\mathtt{C} \in \underline{\mathfrak{B}}(\mathbb{K})$. *On appelle filtre (respectivement idéal) de* $\mathtt{C}$ *et on note* $\uparrow\mathtt{C}$ *(respectivement* $\downarrow\mathtt{C}$*) l'ensemble* $\{\mathtt{D} \in \underline{\mathfrak{B}}(\mathbb{K}) | \mathtt{D} \geq \mathtt{C}\}$ *(respectivement* $\{\mathtt{D} \in \underline{\mathfrak{B}}(\mathbb{K}) | \mathtt{D} \leq \mathtt{C}\}$*)*

1.2.4 Correspondance de Galois

En dehors des travaux de Wille [Wille, 1982], les premières applications de treillis de concepts, notamment par Godin et al. [Godin et al., 1986, Godin et al., 1989], sont issues de découvertes indépendantes faites par Barbut et Monjardet[Barbut and Monjardet, 1970]. Les treillis de concepts ont donc connu à leurs débuts un développement parallèle sous la dénomination de *treillis de Galois*. Cet héritage ce retrouve aujourd'hui dans la littérature et nous faisons ici le lien entre la notion de correspondance de Galois, à l'origine de cette approche, et l'AFC.

Définition 16 (Correspondance de Galois). *Soient* $(\mathtt{E}, \leq_\mathtt{E})$ *et* $(\mathtt{F}, \leq_\mathtt{F})$ *deux ensembles ordonnés. Soient deux fonctions* $\mathtt{f} : \mathtt{E} \to \mathtt{F}$ *et* $\mathtt{g} : \mathtt{F} \to \mathtt{E}$. *le couple* (f, g) *forme une correspondance de Galois si :*

- $\mathtt{f}$ *est antitone :* $\mathtt{x} \leq \mathtt{y} \Rightarrow \mathtt{f}(\mathtt{x}) \geq \mathtt{f}(\mathtt{y})$,
- $\mathtt{g}$ *est antitone,*
- $\mathtt{f} \circ \mathtt{g}$ *est extensive :* $\mathtt{x} \leq \mathtt{f} \circ \mathtt{g}(\mathtt{x})$

– $\mathtt{g} \circ \mathtt{f}$ *est extensive.*

Propriété 5. *Soit* $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$ *un contexte formel. Le couple d'opérateurs de dérivation* $('\,,'\,)$ *définit une correspondance de Galois entre* $(2^{\mathtt{G}}, \subseteq)$ *et* $(2^{\mathtt{M}}, \subseteq)$)

1.2.5 Opérateurs de fermeture

Les propriétés des correspondances de Galois permettent d'établir un lien entre treillis de concepts formels et les systèmes de fermeture[Caspard et al., 2007].

Définition 17 (Opérateur de fermeture)**.** *Soit* $(\mathtt{E}, \leq)$ *un ensemble ordonné. On appelle opérateur de fermeture sur l'ensemble* $\mathtt{E}$ *toute application* $\varphi : \mathtt{E} \to \mathtt{E}$ *qui vérifie les trois propriétés suivantes :*

– *idempotence :* $\forall \mathtt{x} \in \mathtt{E}, \varphi(\varphi(\mathtt{x})) = \varphi(\mathtt{x})$,
– *monotonie :* $\forall (\mathtt{x}, \mathtt{y}) \in \mathtt{E}^2, \mathtt{x} \leq \mathtt{y} \Rightarrow \varphi(\mathtt{x}) \leq \varphi(\mathtt{y})$,
– *extensivité :* $\forall \mathtt{x} \in \mathtt{E}, \mathtt{x} \leq \varphi(\mathtt{x})$.

Propriété 6. *Soit un contexte formel* $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$. *Si on munit* $2^{\mathtt{G}}$ *et* $2^{\mathtt{M}}$ *de l'inclusion ensembliste* $\subseteq$, *les opérateurs de composition* $''$ *de la définition 3 sont des opérateurs de fermeture sur* $(2^{\mathtt{G}}, \subseteq)$ *et* $(2^{\mathtt{M}}, \subseteq)$.

Définition 18 (Fermé)**.** *Étant donné un opérateur de fermeture* φ *sur un ensemble ordonné* $(\mathtt{E}, \leq)$, *un élément* $\mathtt{x}$ *de* $\mathtt{E}$ *est dit « fermé » si* $\mathtt{x} = \varphi(\mathtt{x})$.

Propriété 7. *Soient* $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$ *un contexte formel,* $\mathfrak{B}(\mathbb{K})$ *l'ensemble de ses concepts formels,* $\mathfrak{U}(\mathbb{K})$ *l'ensemble des extensions de ces concepts et* $\mathfrak{J}(\mathbb{K})$ *l'ensemble de leurs intensions.*

– $\mathfrak{U}(\mathbb{K})$ *est l'ensemble des fermés de* $2^{\mathtt{G}}$ *par l'opérateur de composition* $''$.
– $\mathfrak{J}(\mathbb{K})$ *est l'ensemble des fermés de* $2^{\mathtt{M}}$ *par l'opérateur de composition* $''$.

1.3 Visualisation de treillis

L'un des intérêts majeurs de l'AFC repose sur les capacités de représentation visuelle des treillis de concepts dans un diagramme appelé *diagramme de Hasse.*

1.3.1 Diagramme de Hasse d'un ensemble ordonné

En mathématiques, le diagramme de Hasse, du nom du mathématicien allemand Helmut Hasse, est une représentation visuelle d'un ordre fini. Similaire à la représentation habituelle d'un graphe sur papier, il en facilite la compréhension. Soit $(\mathtt{E}, \leq)$ un ensemble ordonné fini. Pour dessiner un diagramme de Hasse :

- On représente les éléments de l'ordre par des points.
- Si un élément $\mathtt{x}$ est plus grand qu'un autre élément $\mathtt{y}$ selon $\leq$, on place la représentation de $\mathtt{x}$ plus haut que celle de $\mathtt{y}$.
- Deux éléments en relation sont reliés par un arc. Du fait de la disposition des points, il est inutile que cet arc soit orienté.
- Pour ne pas surcharger le schéma, on ne représente pas toute la relation d'ordre, mais seulement sa réduction transitive. Soient $(\mathtt{x}, \mathtt{y}) \in \mathtt{E}^2, \mathtt{x} < \mathtt{y}$. S'il existe un élément $\mathtt{z} \in \mathtt{E}$ tel que $\mathtt{x} < \mathtt{z} < \mathtt{y}$, alors l'arc $(\mathtt{x}, \mathtt{y})$ n'est pas représenté. D'autre part on ne représente pas les boucles d'un élément vers lui-même.
- On veille autant que possible à ne pas croiser les segments.

Soit $\mathtt{E} = \{1, 2, 3, 5, 6, 10, 12, 15, 30\}$ l'ensemble des diviseurs de 30, muni de la relation d'ordre $\div$ « divise ». La figure 1.3.1 montre le diagramme de Hasse de $(\mathtt{E}, \div)$. Sur cette figure, on peut lire que 3 divise 6 et 15, qui eux mêmes divisent 30. La relation « 3 divise 30 » est lue par transitivité.

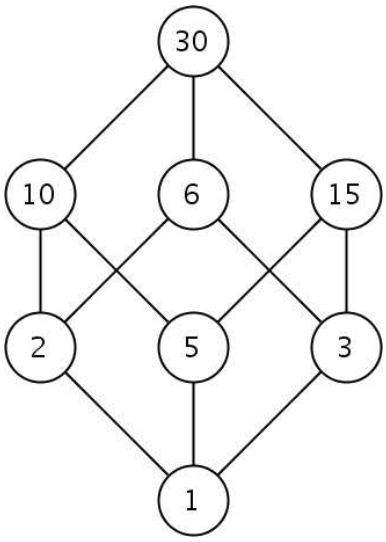

FIG. 1.1 – Diagramme de Hasse des diviseurs de 30.

1.3.2 Diagramme de Hasse d'un treillis de concepts

Un treillis de concepts peut être visualisé grâce à un diagramme de Hasse. Les nœuds représentent des concepts et les arcs matérialisent la relation sous-concept/super-concept. La figure 1.2 montre le contexte formel `pathologie` et son treillis de concepts. Chaque concept est étiqueté au dessus par son intension et en dessous par son extension.

	diabete	hypertension	dyslipidemie	angor
p1	×	×		
p2		×		
p3		×		
p4	×			
p5	×		×	×
p6	×			
p7	×	×	×	×
p8	×	×		

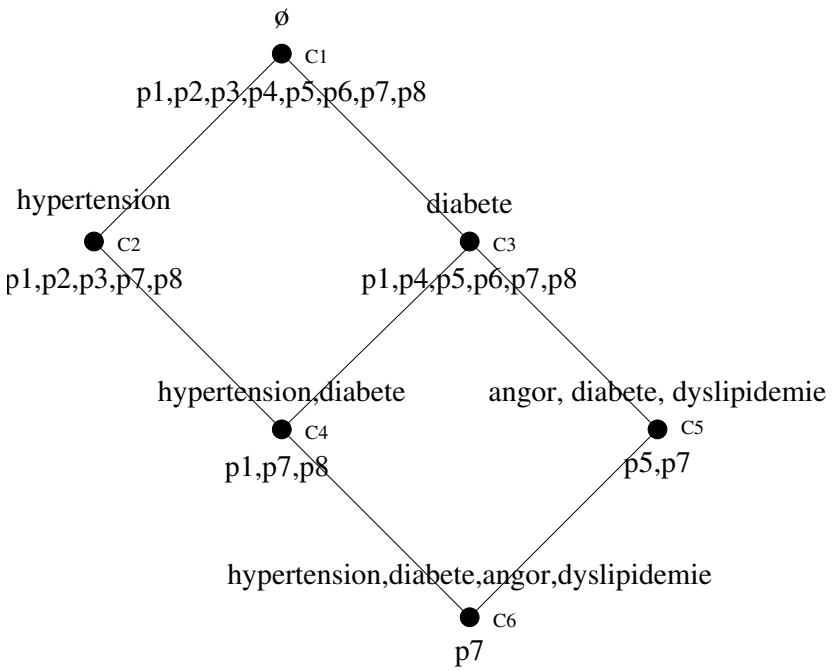

FIG. 1.2 – Le contexte formel `pathologie` et son treillis de concepts.

1.3.3 Simplifications du diagramme de Hasse

La structure d'un treillis de concepts comporte beaucoup de redondances. Pour alléger le diagramme de Hasse et faciliter la lecture, il est possible de représenter un treillis de concepts sous une forme plus compacte, sans perte d'information, sous la forme de treillis d'héritage [Godin et al., 1995a].

Treillis d'héritage

Dans un treillis de concepts, si un objet est présent dans l'extension d'un concept, il apparaît également dans l'extension de tous ses super-concepts. Sur un diagramme de Hasse, il est possible de ne représenter cet objet que dans l'infimum des concepts où il apparaît. Sur la figure 1.2, `p1` apparaît dans les concepts `C1`, `C2`, `C3`, `C4`. `C4` est le plus petit concept dont l'extension contient `p1`. On peut éliminer `p1` de l'extension de `C1`, `C2`, `C3` et retrouver cette information par héritage en parcourant le filtre de `C4`. De manière duale, un attribut présent dans l'intension d'un concept apparaît également dans l'intension de tous les concepts de son idéal. Il est possible de supprimer cette redondance en ne conservant l'attribut que dans l'intension du supremum des concepts où il apparaît.

Formellement on peut définir deux types de treillis d'héritage, selon les objets, ou selon les attributs.

Définition 19 (Treillis d'héritage). *Soit* $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$ *un contexte formel et* $\underline{\mathfrak{B}}(\mathbb{K})$ *son treillis de concepts. On appelle treillis d'héritage selon les objets de* $\mathbb{K}$*, l'ensemble des couples* $(\mathtt{X}, \mathtt{Y}) \in 2^{\mathtt{G}} \times 2^{\mathtt{M}}$ *tels que :*

- $\mathtt{Y} = \mathtt{Y}''$
- $\mathtt{X}$ *est l'ensemble des éléments de* $\mathtt{Y}'$ *qui n'apparaissent dans aucune des extensions des concepts* $\mathtt{C} \in \underline{\mathfrak{B}}(\mathbb{K}), \mathtt{C} > (\mathtt{Y}', \mathtt{Y})$

Les éléments de $\mathtt{X}$ *sont appelés objets propres du concept* $(\mathtt{Y}', \mathtt{Y})$.

De manière duale on peut définir un treillis d'héritage selon les attributs de $\mathbb{K}$ *comme l'ensemble des couples* $(\mathtt{X}, \mathtt{Y}) \in 2^{\mathtt{G}} \times 2^{\mathtt{M}}$ *tels que :*

- $\mathtt{X} = \mathtt{X}''$
- $\mathtt{Y}$ *est l'ensemble des éléments de* $\mathtt{X}'$ *qui n'apparaissent dans aucune des intensions des concepts* $\mathtt{C} \in \underline{\mathfrak{B}}(\mathbb{K}), \mathtt{C} < (\mathtt{X}, \mathtt{X}')$

Les éléments de $\mathtt{Y}$ *sont appelés attributs propres du concept* $(\mathtt{X}, \mathtt{X}')$.

À partir du contexte `pathologie` (tableau 1.1), on obtient les treillis d'héritage de la figure 1.3.

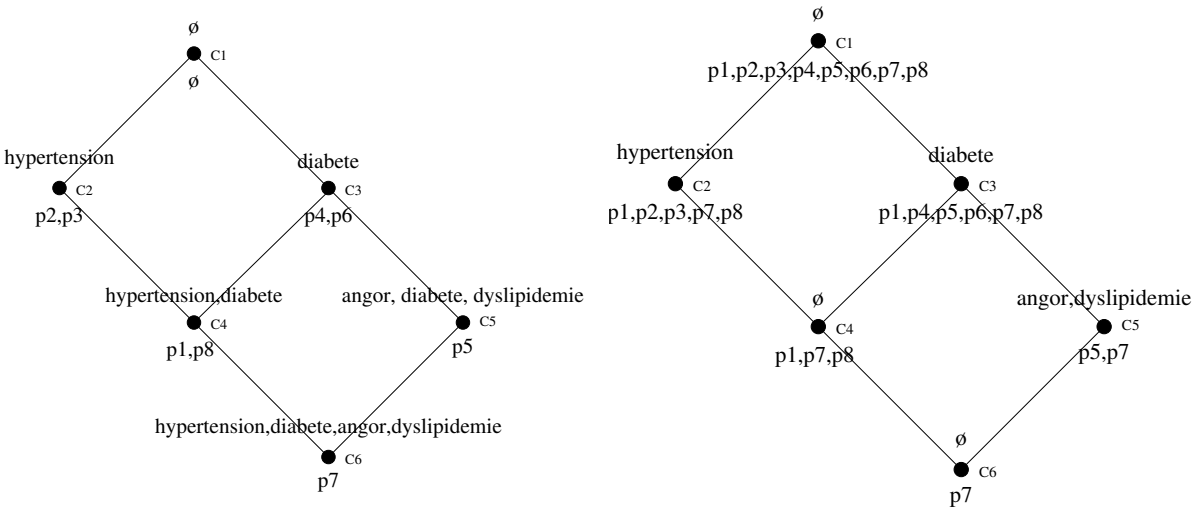

FIG. 1.3 – Treillis d'héritage selon les objets (gauche) et les attributs (droite) du contexte `pathologie`.

Les deux types de treillis d'héritage peuvent être couplés pour obtenir un treillis d'héritage complet selon les objets et les attributs. On parle encore d'étiquetage réduit[1], dont un exemple est donné sur la figure 1.4.

[1] à l'opposé, on parlera d'étiquetage complet

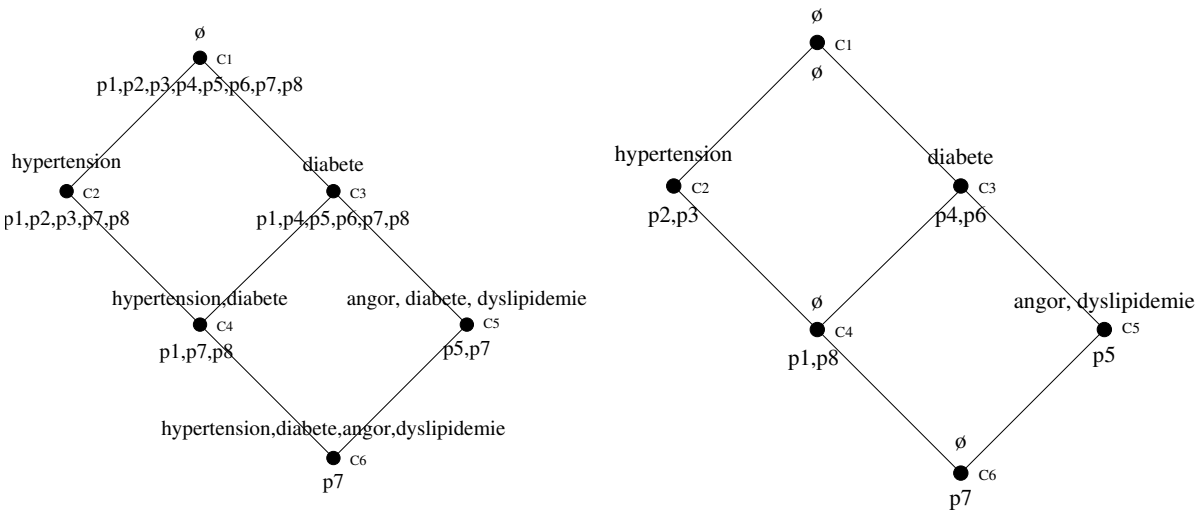

FIG. 1.4 – Le treillis du contexte `pathologie` : à gauche, étiquetage complet, à droite, étiquetage réduit.

1.4 Extensions de l'AFC

Par définition, l'AFC traite des contextes binaires, ce qui limite l'exploration de données représentées sous une forme plus complexe. En particulier, L'AFC ne permet pas au départ de manipuler des attributs numériques, ou gérer des modèles plus complexes comme les modèles relationnels. Ces problèmes ont donné lieu à des développements de l'AFC qui prennent en compte des structures de données plus élaborées.

1.4.1 Contextes multivalués

Dans de nombreuses situations, la description d'un objet nécessite, pour être plus expressive, l'utilisation d'attributs non binaires. En AFC de tels objets sont modélisés par des attributs multivalués [Ganter and Wille, 1999].

Attributs multivalués

Définition 20 (Contexte multivalué)**.** *Un contexte multivalué est un quadruplet* $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{W}, \mathtt{I})$ *où* $\mathtt{G}$ *et* $\mathtt{M}$ *sont des ensembles d'objets et d'attributs respectivement,* $\mathtt{W}$ *un ensemble de valeurs, et* $\mathtt{I} \subseteq \mathtt{G} \times \mathtt{M} \times \mathtt{W}$ *une relation ternaire dans laquelle un tuple* $(\mathtt{g}, \mathtt{m}, \mathtt{w})$ *signifie « l'objet* $\mathtt{g}$ *a la valeur* $\mathtt{w}$ *pour l'attribut* $\mathtt{m}$ *».*

Les éléments de $\mathtt{M}$ *sont appelés attributs multivalués. Chaque attribut multivalué peut être vu comme une application partielle de* $\mathtt{G}$ *dans* $\mathtt{M}$ *et on notera* $\mathtt{m}(\mathtt{g}) = \mathtt{w}$ *si* $(\mathtt{g}, \mathtt{m}, \mathtt{w}) \in \mathtt{I}$*. On définit*

alors le domaine d'un attribut multivalué $\mathtt{m} \in \mathtt{M}$ *comme l'ensemble*

$$\mathtt{dom(m)} = \{\mathtt{g} \in \mathtt{G} | \exists \mathtt{w} \in \mathtt{W}, \mathtt{m(g)} = \mathtt{w}\}$$

Le tableau 1.3 montre le contexte multivalué **cancer du sein** indiquant pour un ensemble de tumeurs (t1, ..., t7) leur taille en centimètres et la présence de ganglions dans certains territoires (**aucune**, **axillaire**, **autre**).

	taille (cm)	atteinte ganglionnaire
t1	1.0	aucune
t2	3.2	axillaire
t3	0.5	aucune
t4	5.1	autre
t5	4.7	axillaire
t6	2.6	aucune
t7	4.1	axillaire

TAB. 1.3 – Contexte multivalué **cancer du sein**.

Pour pouvoir traiter un contexte multivalué, celui-ci doit être transformé en un contexte dérivé monovalué (binaire) équivalent. C'est le rôle des méthodes de «scaling» ou mise à l'échelle.

Échelle conceptuelle

Il n'y a pas de méthode unique pour dériver un contexte multivalué en contexte monovalué. Les résultats obtenus dépendent des choix qui conduisent à la transformation de chaque attribut du contexte multivalué en un contexte formel appelé *échelle conceptuelle* : c'est le processus de *scaling conceptuel.*

Définition 21 (Échelle conceptuelle)**.** *Une échelle conceptuelle pour un attribut* $\mathtt{m}$ *d'un contexte multivalué donné est un contexte (monovalué)* $\mathbb{S}_\mathtt{m} = (\mathtt{G_m}, \mathtt{M_m}, \mathtt{V_m})$ *où* $\mathtt{m(G)} \subseteq \mathtt{G_m}$*. Les objets de l'échelle sont appelés valeurs d'échelle alors que ses attributs sont appelés attributs d'échelle.*

En théorie, n'importe quel contexte peut servir d'échelle : nominal, ordinal, biordinal ou dichotomique [Ganter and Wille, 1999]. Par exemple, l'attribut **atteinte ganglionnaire** du contexte **cancer du sein** peut être interprété à l'aide d'une échelle nominale en considérant que ses valeurs sont mutuellement exclusives.

Pour la taille de la tumeur, on peut utiliser l'échelle ordinale représentée sur la figure 1.6.

Il n'y a pas de méthode univoque pour définir une échelle conceptuelle. Le choix de celle-ci est généralement dicté par un expert du domaine, plus à même d'identifier les transformations les plus pertinentes.

Atteinte ganglionnaire	aucune	axillaire	autre
aucune	×		
axillaire		×	
autre			×

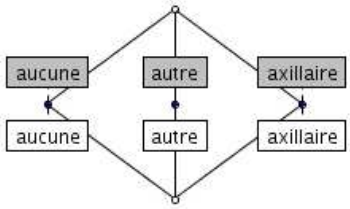

FIG. 1.5 – Échelle nominale de l'attribut atteinte axillaire.

taille	≥ 2	≥ 5
0.5	×	×
1.0	×	×
2.6		×
3.2		×
4.1		×
4.7		×
5.1		

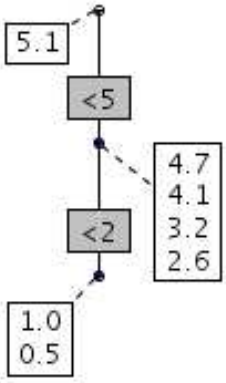

FIG. 1.6 – Échelle ordinale de l'attribut taille.

La dernière étape du scaling consiste à substituer dans le contexte multivalué les attributs d'échelle aux attributs multivalués. En pratique, les objets du contexte multivalué sont inchangés, et chaque attribut multivalué est remplacé par la ligne qui lui correspond dans l'échelle. Ainsi, pour le contexte multivalué **cancer du sein**, on obtient le contexte dérivé représenté dans la figure 1.7.

	taille			atteinte ganglionnaire		
	≥ 2	≥ 5	≥ 5	aucune	axillaire	autre
t1	×	×		×		
t2	×	×			×	
t3	×	×		×		
t4			×			×
t5		×			×	
t6		×		×		
t7		×			×	

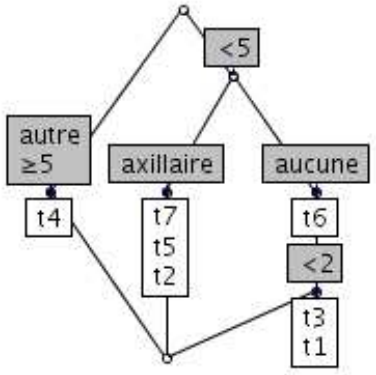

FIG. 1.7 – Le contexte dérivé du contexte multivalué **cancer du sein** et son treillis.

1.4.2 Diagrammes de Hasse entrelacés

L'analyse de contextes multivalués peut bénéficier d'une forme de représentation particulière des treillis de concepts : les diagrammes de Hasse entrelacés [Wille, 1989, Vogt et al., 1991]. Ce type de diagramme repose sur une partition des attributs formels. Soit $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$ un contexte formel et $\{\mathtt{M}_1, \mathtt{M}_2\}$ une partition de $\mathtt{M}$. Le principe consiste à construire le diagramme de Hasse du treillis $\underline{\mathfrak{B}_1}(\mathtt{G}, \mathtt{M}_1, \mathtt{I} \cap \mathtt{G} \times \mathtt{M}_1)$ et à remplacer chacun de ses nœuds par une copie du treillis $\underline{\mathfrak{B}_2}(\mathtt{G}, \mathtt{M}_2, \mathtt{I} \cap \mathtt{G} \times \mathtt{M}_2)$. Ainsi, en partitionnant le contexte **cancer du sein** selon les échelles de l'attribut **taille** et de l'attribut **atteinte ganglionnaire**, on obtient les treillis entrelacés de la figure 1.8.

1.4.3 Analyse Relationnelle de Concepts

Alors que l'AFC classique est une méthode de choix pour explorer les relations entre des individus et une conjonction d'attributs binaires ou multivalués, celle-ci échoue à tirer parti de propriétés qui relieraient les individus entre eux. C'est ce type de relation que l'on rencontre

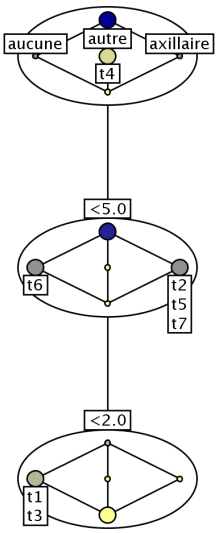

FIG. 1.8 – Treillis entrelacés du contexte **cancer du sein** : l'attribut **atteinte ganglionnaire** est imbriqué dans l'attribut **taille**. Chaque nœud du treillis de l'échelle de l'attribut **taille** contient une copie du treillis de l'échelle de l'attribut **atteinte ganglionnaire**.

dans les modèles *Entité-Relation* (ER) utilisés dans le domaine des bases de données relationnelles. On peut illustrer cette situation par une base de données modélisant des chercheurs, des thèmes de recherche, des publications et des références entre auteurs et publications. Ce type de modèle induit des liens entre publications et entre auteurs par l'intermédiaire de la relation de *citation* que l'on peut qualifier d'attribut relationnel. Pour traiter les attributs relationnels, [Hacene et al., 2007b] proposent l'Analyse Relationnelle de Concepts (ARC). L'ARC produit un ensemble de treillis de concepts à partir d'une collection de contextes formels et de relations entre ces contextes. Les concepts des différents treillis obtenus sont eux mêmes associés par des relations qui sont stockées dans un graphe auxiliaire.

Formellement, l'ARC s'appuie sur la structure de données de Famille Relationnelle de Contextes définie comme suit :

Définition 22 (Famille Relationnelle de Contextes)**.** *Une famille relationnelle de contextes est une paire* $(\mathtt{K}, \mathtt{R})$ *où* $\mathtt{K}$ *est un ensemble de contextes formels* $\mathbb{K}_\mathtt{i} = (\mathtt{G_i}, \mathtt{M_i}, \mathtt{I_i})$*,* $\mathtt{R}$ *un ensemble de relations* $\mathtt{r_k} \subseteq \mathtt{G_i} \times \mathtt{G_j}$ *où* $\mathtt{G_i}$ *et* $\mathtt{G_j}$ *sont les ensembles d'objets des contextes* $\mathbb{K}_\mathtt{i}$ *et* $\mathbb{K}_\mathtt{j}$*.*

Pour construire les treillis de concepts relationnels, l'ARC met en œuvre un mécanisme de scaling relationnel proche du scaling conceptuel. Le scaling relationnel élève les relations de $\mathtt{R}$ au rang d'attributs multivalués selon le principe suivant : soient $\mathbb{K}_\mathtt{i} = (\mathtt{G_i}, \mathtt{M_i}, \mathtt{I_i})$ et $\mathbb{K}_\mathtt{j} = (\mathtt{G_j}, \mathtt{M_j}, \mathtt{I_j})$ deux contextes formels et une relation $\mathtt{r} \subseteq \mathtt{G_i} \times \mathtt{G_j}$. Le scaling relationnel va

enrichir $\mathtt{M_i}$ et $\mathtt{I_i}$ par des concepts $\mathtt{c_j}$ du treillis $\underline{\mathfrak{B}_j}(\mathbb{K}_j)$: un objet $\mathtt{g_i} \in \mathtt{G_i}$ sera associé à des concepts $\mathtt{c_j}$ si leur extension contient tout où partie des éléments de $\mathtt{r}(\mathtt{g_i})$[2]. Le contexte $\mathbb{K}_\mathtt{i}$ ainsi étendu va donner naissance à un nouveau treillis de concepts. Ce treillis fait apparaître des relations entre les objets de $\mathtt{G_i}$ et, non plus des objets de $\mathtt{G_j}$, mais leur abstraction par des concepts de $\underline{\mathfrak{B}_j}(\mathbb{K}_j)$. Le nombre de relations d'une famille relationnelle de contextes et la possibilité de dépendances circulaires entre objets nécessite la mise en œuvre d'un mécanisme itératif de construction de treillis appelée MULTI-AFC détaillé dans [Hacene et al., 2007b].

L'ARC a été appliquée avec succès dans la construction d'ontologies à partir de textes [Bendaoud et al., 2007a] ou encore la ré-ingénierie des modèles UML [Hacene et al., 2007a].

1.5 Algorithmes de construction de treillis

Le problème de génération de l'ensemble des concepts formels d'un treillis et de son diagramme de Hasse a fait l'objet de nombreuses études. Le nombre de concepts d'un treillis peut être une fonction exponentielle de la taille de son contexte. Kusnetsov [Kuznetsov, 1989] a montré que déterminer ce nombre est un problème #P-complet. Cette classe regroupe les problèmes d'énumération de solutions à des problèmes de classe NP. Les problèmes de la classe #P posent la question « Combien y a-t-il ? » plutôt que « Y a-t-il ? ». Par exemple : « Combien y a-t-il de sous-ensembles d'une liste d'entiers dont la somme est égale à zéro ? ». Guénoche [Guénoche, 1990] propose une revue des premiers algorithmes de construction de treillis, complétée par celle de Kuznetsov et Obiedkov [Kuznetsov and Obiedkov, 2001].

Ces algorithmes peuvent être classés en trois catégories :

- les algorithmes *batchs*,
- les algorithmes incrémentaux,
- les algorithmes d'assemblage.

Les algorithmes batchs construisent le treillis à partir d'un contexte formel complet. L'évolution des données (ajout d'objet/attribut au contexte) entraîne à nouveau la construction du treillis. Parmi ces algorithmes, on peut citer ceux de Chein [Chein, 1969], Bordat [Bordat, 1986], Ganter [Ganter, 1984, Ganter and Wille, 1999] ou AI-tree [Zabezhailo et al., 1987]. Les algorithmes incrémentaux remédient au problème de la reconstruction du treillis dans le cadre de contextes dynamiques. Ces algorithmes effectuent des mises à jour locales du treillis après l'ajout d'un objet dans le contexte formel. Parmi les algorithmes incrémentaux, on trouve ceux de Nourine [Nourine and Raynaud, 1999], Norris [Norris, 1978] et Godin [Godin et al., 1995b]. Enfin, les algorithmes d'assemblage, comme celui de Valtchev [Valtchev et al., 2002], généralisent le caractère incrémental à des ensembles d'objets/attributs. Ces algorithmes se basent

[2]Il existe différentes méthodes d'association décrites en détail dans [Hacene et al., 2007b]

sur la fusion de deux contextes et les produits de leurs treillis respectifs. Cette technique peut également être utilisée comme une stratégie « diviser pour régner » de construction de treillis dans un contexte statique.

Kuznetsov et Obiedkov [Kuznetsov and Obiedkov, 2001] ont étudié la complexité de la plupart de ces algorithmes, en apportant quelques optimisations pour certains. Pour un contexte $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$ et son treillis de concepts $\mathfrak{B}$, le tableau 1.4 donne leur complexité en temps, au pire cas, et pour les algorithmes batch, leur délai polynomial.

Algorithme	Complexité en temps	Délai polynomial
Bordat	$\mathcal{O}(\lvert\mathtt{G}\rvert\lvert\mathtt{M}\rvert^2\lvert\mathfrak{B}\rvert)$	$\mathcal{O}(\lvert\mathtt{G}\rvert\lvert\mathtt{M}\rvert^2)$
Ganter	$\mathcal{O}(\lvert\mathtt{G}\rvert^2\lvert\mathtt{M}\rvert\lvert\mathfrak{B}\rvert)$	$\mathcal{O}(\lvert\mathtt{G}\rvert^2\lvert\mathtt{M}\rvert)$
Close By One	$\mathcal{O}(\lvert\mathtt{G}\rvert^2\lvert\mathtt{M}\rvert\lvert\mathfrak{B}\rvert)$	$\mathcal{O}(\lvert\mathtt{G}\rvert^3\lvert\mathtt{M}\rvert)$
Nourine	$\mathcal{O}(\lvert\mathtt{G}\rvert + \lvert\mathtt{M}\rvert)\lvert\mathtt{G}\rvert\lvert\mathfrak{B}\rvert)$	
Godin	$\mathcal{O}(\lvert\mathfrak{B}\rvert^2)$	

TAB. 1.4 – Complexité des principaux algorithmes de construction d'un treillis $\mathfrak{B}$ à partir de son contexte $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$.

La complexité en temps au pire cas de l'algorithme d'assemblage de Valtchev admet la borne $\mathcal{O}((\mathtt{k}+\mathtt{m})(\mathtt{b_1 b_2} + \mathtt{bm}))$ où $\mathtt{k}$ est le nombre total d'objets du contexte assemblé, $\mathtt{m}$ son nombre d'attributs, $\mathtt{b_i}$ la taille des treillis avant assemblage et $\mathtt{b}$ la taille du treillis final.

1.6 Logiciels de manipulation de treillis

Plusieurs programmes ont été développés pour construire, manipuler et visualiser des treillis de concepts. Un des plus anciens est ConImp [3] disponible sous DOS et Linux, il fonctionne en mode texte et ne permet pas de visualiser les treillis. Parmi les logiciels spécifiques les plus récents, on distingue Toscana [Vogt and Wille, 1994], Concept Explorer [Yevtushenko, 2000] et Galicia [Valtchev et al., 2003] tous trois distribués en licence open-source. Toscana est actuellement développé en Java (ToscanaJ) par des équipes des universités de Darmstadt et du Queensland [4]. L'une des particularités de ToscanaJ est de construire et visualiser des treillis entrelacés (Fig. 1.9).

Dans nos travaux, nous avons principalement utilisé Conexp [5] (Fig. 1.10) et Galicia [6]. Ces applications sont dotées de toutes les fonctionnalités de base pour manipuler des treillis (édition de contexte formel, construction et visualisation). De plus, Galicia présente certaines

[3] `http://www.mathematik.tu-darmstadt.de/~burmeister/`
[4] `http://toscanaj.sourceforge.net/`
[5] `http://sourceforge.net/projects/conexp`
[6] `http://www.iro.umontreal.ca/~galicia/`

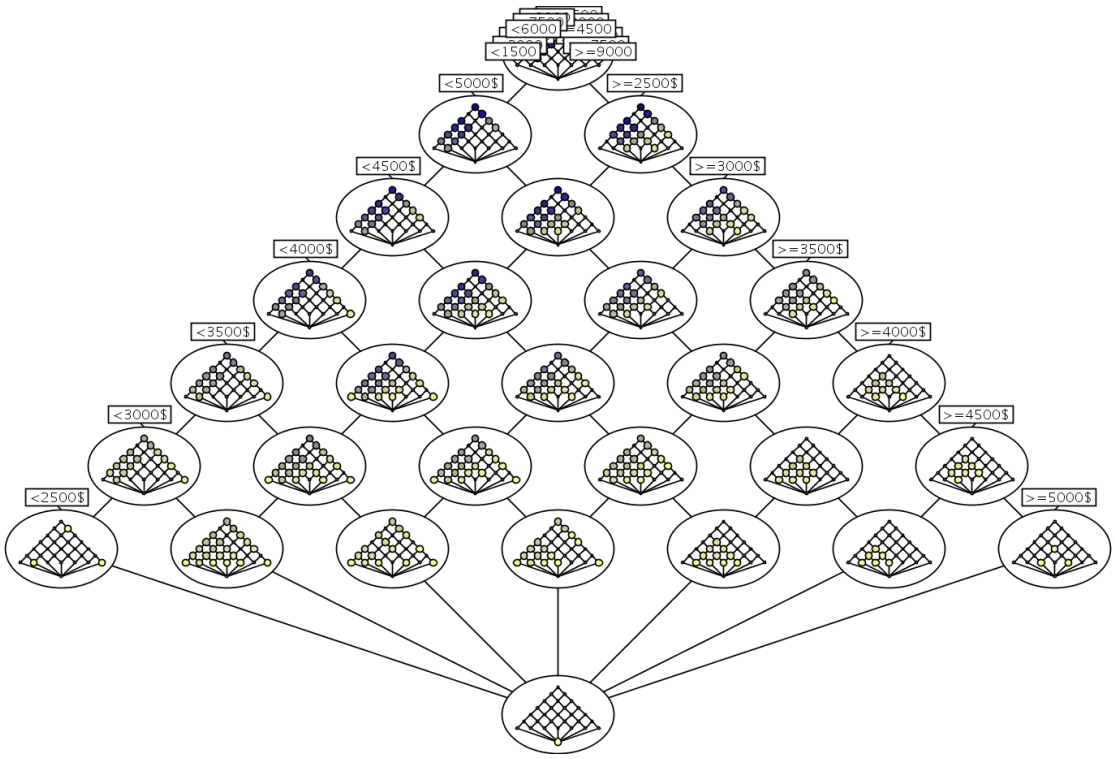

FIG. 1.9 – Des treillis entrelacés dans ToscanaJ.

fonctionnalités avancées comme la manipulation de contextes relationnels, la fusion de treillis ou encore la construction d'icebergs de concepts (Fig. 1.11).

D'autres applications, non spécifiques de l'AFC, manipulent des treillis de concepts. On peut citer de manière non exhaustive Coron [Szathmary, 2006], une plateforme de fouille de données, HierMail [Eklund and II, 2002], un système de gestion de courriel basé sur l'AFC, ou encore Camelis [Ferré, 2002], un système d'information logique qui permet d'indexer et classer des documents numériques puis de naviguer dans le treillis de concepts construit à partir de leurs propriétés.

1.7 Applications de l'analyse formelle de concepts

L'AFC a donné lieu à de nombreuses applications dans des domaines aussi divers que l'extraction des connaissances à partir de données, le raisonnement et la représentation des connaissances, la recherche d'information, la linguistique ou encore le génie logiciel. Son succès tient peut-être aux liens qu'elle établit entre le formalisme mathématique et les modes de représentation et le raisonnement humain. Avant de faire un état des lieux de ses principales applications, nous donnons une courte interprétation philosophique de l'AFC.

1.7.1 Interprétation philosophique de l'AFC

En philosophie, un concept est une « *représentation abstraite et générale d'une chose ou d'un fait* »[7]. Pour Bergson : « Le philosophe, remontant du percept au concept, voit se condenser en logique tout ce que le physique avait de réalité positive »[Bergson, 1907]. Selon Kant : « [un concept] est un moyen de juger, c'est-à-dire un universel, une forme de classe sous laquelle on peut subsumer un singulier » [Hamelin, 1925]. Pour les logiciens, il se compose d'une extension (ou dénotation) et d'une intension (ou compréhension). Le concept peut être ou le sujet, ou l'attribut, dans une infinité de jugements possibles. Soit le concept d'Homme : son extension est la proposition prédicative : « un tel (ou un tel) est un Homme (exemple : Socrate est un Homme), et sa compréhension est la proposition réflexive : « l'Homme est tel (et tel) » (exemple : l'Homme est raisonnable, social, animal, bipède, mammifère, vertébré, etc.). Dans le premier cas, Homme est attribut ou prédicat et définit une classe ; dans le second cas, comme sujet du jugement, il détermine l'ensemble des caractères qui le définissent. Notons, en outre, que l'extension d'un concept est toujours inversement proportionnelle à sa compréhension. « *Il y a des concepts plus généraux que d'autres. Les concepts les plus généraux, celui d'animal par exemple, ont un petit nombre de propriétés mais ils ont une grande extension. Les concepts plus spécifiques, comme caniche par exemple, ont un plus grand nombre de propriétés et en*

[7] http ://fr.encarta.msn.com/encyclopedia_741534111/concept_(philosophie).html

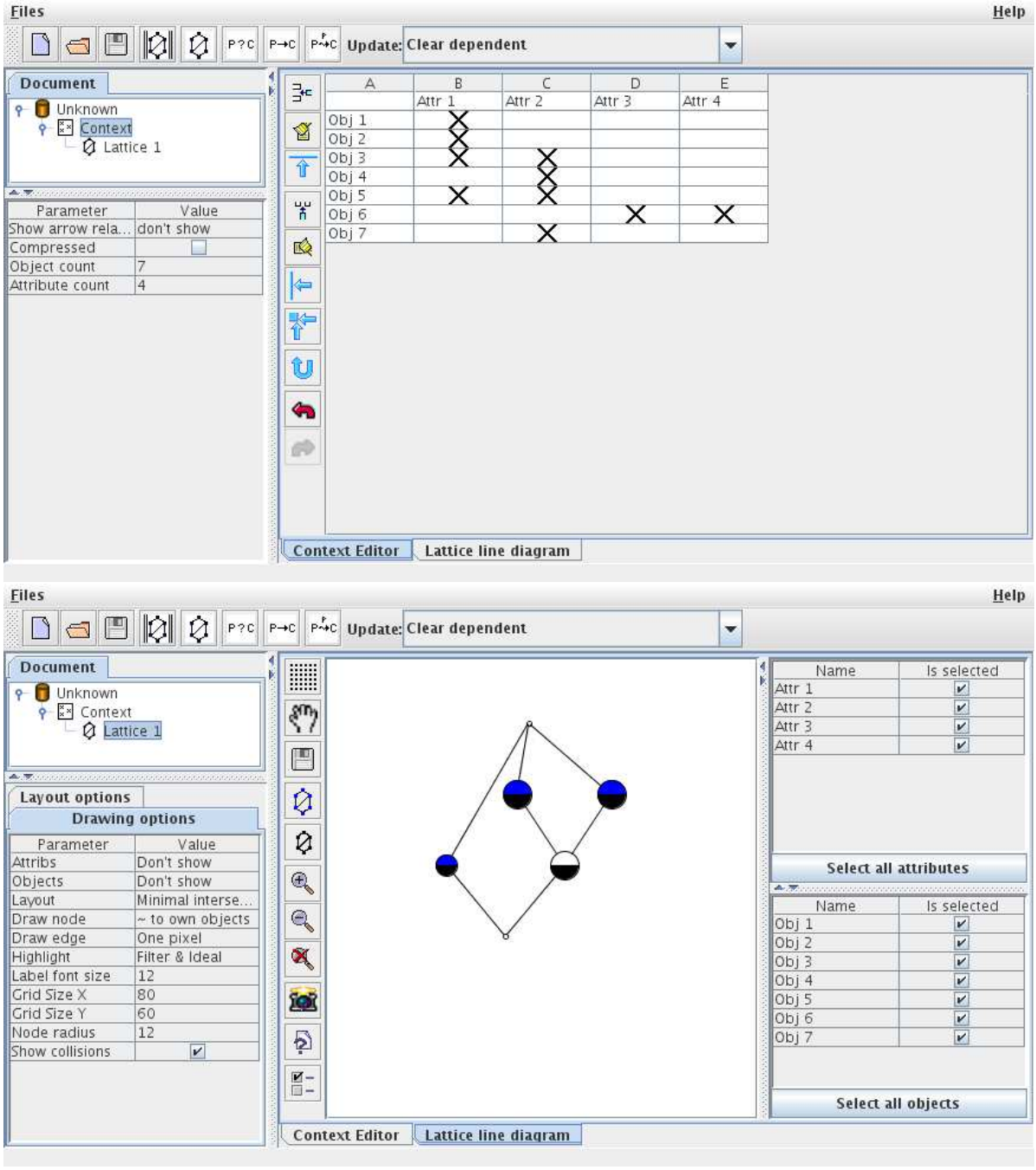

FIG. 1.10 – Conexp : édition d'un contexte formel et son treillis correspondant.

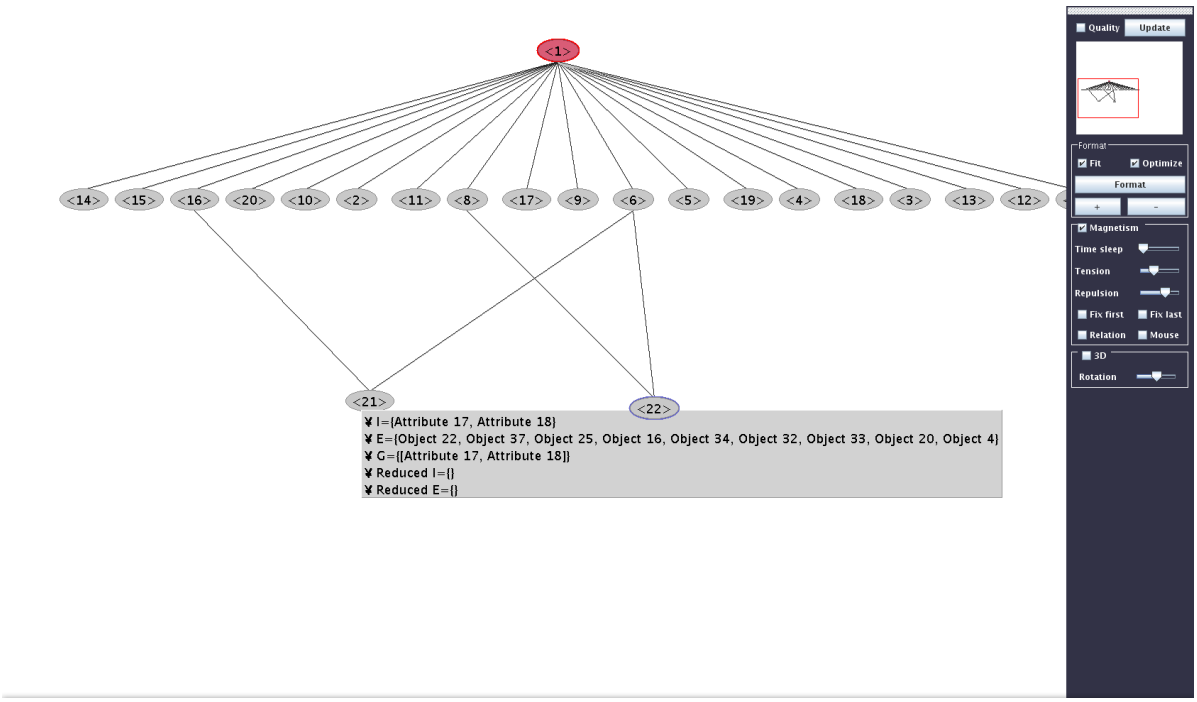

FIG. 1.11 – Un iceberg construit par Galicia.

conséquence les classes qu'ils définissent sont moins étendues. La richesse de la définition (la compréhension) est en relation inverse avec l'extension. » [Richard and Richard, 1995]. Cette définition philosophique d'un concept illustre l'intérêt de l'approche par AFC.

Wille [Wille, 2005] va plus loin en rapprochant l'AFC de la théorie philosophique de Seiler qui considère les concepts comme des structures cognitives dont la formation dans l'esprit humain est un processus constructif et adaptatif. L'article établit un parallèle sur les douze aspects qui caractérisent les concepts philosophiques dans la théorie de Seiler. Nous en citons deux exemples à titre illustratif :

- *les concepts sont des actes cognitifs et des unités de connaissance potentiellement indépendants du langage.* Les concepts formels peuvent mathématiser les concepts philosophiques comme unités de connaissance représentées par leur intension, ou leur extension, indépendamment du langage. En outre, la construction du diagramme de Hasse d'un treillis de concepts stimule l'acte de cognition et la représentation personnelle ou conventionnelle des concepts philosophiques.
- *Les concepts sont spécifiques d'un domaine et souvent prototypiques.* Pour Seidel, les concepts personnels sont issus de l'expérience et de conceptions implicites, ce qui limite leur portée et leur validité. La nature implicite des concepts philosophiques confère un caractère prototypique à certains faits qu'ils sont censés représenter. Les concepts formels

dérivent de contextes formels qui sont une représentation partielle d'un domaine d'intérêt donné. En cela, ils sont également limités dans leur portée. Ils peuvent néanmoins activer des conceptions implicites de ce domaine et illustrer leur nature prototypique. Sur ce dernier point, Wille fait référence à la théorie des protoconcepts [Wille, 2000]. Étant donné un contexte formel $(\mathtt{G}, \mathtt{M}, \mathtt{I})$, un protoconcept est un couple $(\mathtt{A}, \mathtt{B}) \in \mathtt{G} \times \mathtt{M}$ vérifiant $\mathtt{A}'' = \mathtt{B}'$. Dans un tel couple, les objets de $\mathtt{A}$ sont prototypiques de tous les objets partageant les attributs de $\mathtt{B}$.

Wille souligne également l'intérêt de l'AFC comme support pragmatique du raisonnement [Wille, 1999]. Épistémologiquement, le raisonnement est un processus illimité d'investigation. A chaque étape de ce processus, les investigateurs traitent de problèmes nécessairement restreints, qui reposent grandement sur le sens commun, des points de vue et des objectifs liés à la réalité en question. C'est pourquoi le savoir humain est toujours incomplet et fait continuellement appel à la communication inter-subjective et l'argumentation pour se développer. Ce dernier aspect fait de l'AFC un outil de l'*action communicative* théorisée par Habermas [Habermas, 1981].

Enfin, Wille utilise la métaphore de *paysage* dans lequel les explorateurs commencent par découvrir leur environnement immédiat puis lancent des expéditions de plus en plus lointaines. Il présente l'AFC comme un paradigme de navigation conceptuel dans le paysage du savoir, capable de supporter de nombreuses tâches du traitement des connaissances : exploration, recherche, reconnaissance, identification, analyse, investigation, décision, amélioration, restructuration et mémorisation.

1.7.2 Extraction des connaissances à partir de données

L'extraction de connaissances à partir de données (ECD) est le processus d'identification de motifs de connaissances nouveaux, intelligibles et potentiellement utiles à partir de données [Fayyad et al., 1996]. L'ECD est un processus cyclique qui regroupe plusieurs étapes :

- une étape de sélection de données cibles dans un entrepôt de données,
- une étape de pré-traitement,
- une étape de transformation,
- une étape de fouille de données ou data mining,
- une étape de visualisation/interprétation des résultats.

À la dernière étape du processus, les connaissances nouvelles peuvent servir de point de départ à un nouveau cycle (figure 1.12).

L'étape de fouille de données, considérée souvent comme le cœur de l'ECD, consiste à découvrir des régularités dans les données. La conception de systèmes capables de traiter efficacement de gros volumes de données, de tenir compte d'informations manquantes ou erronées et d'of-

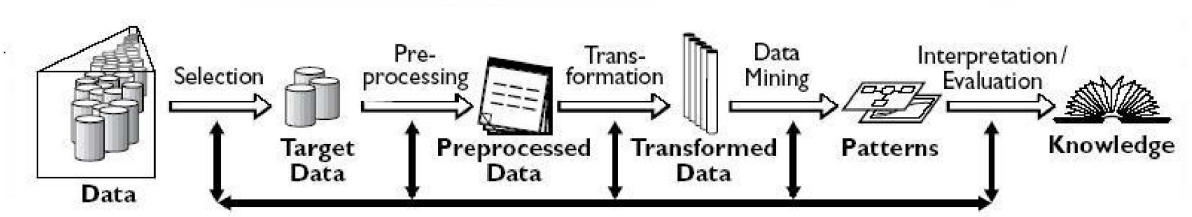

FIG. 1.12 – Le processus d'extraction de connaissances à partir de données (extrait de [Fayyad et al., 1996]).

frir des mécanismes de visualisation des résultats a motivé l'essentiel des recherche depuis les années 1990. L'AFC, en se fondant sur la construction de hiérarchies conceptuelles à partir d'un ensemble d'observations, et en permettant leur visualisation dans un treillis de concepts, est donc une méthode d'ECD par excellence [Valtchev et al., 2004]. C'est en particulier dans le domaine des règles d'association que l'on a assisté à l'émergence des principales applications de l'AFC en extraction des connaissances [Pasquier et al., 1999c, Zaki and Hsiao, 2002, Stumme et al., 2002]. La propriété de fermeture de l'AFC est d'ailleurs une des clefs de ce succès : elle permet de n'extraire que des motifs de taille maximale et réduit considérablement l'effort d'exploration et d'interprétation de l'analyste. Mais le potentiel de l'AFC dans l'ECD va au-delà et pourrait permettre de résoudre de nombreux problèmes posés en situation réelle, comme le traitement de données dynamiques, distribuées, stockées dans des formats riches ou de structure particulière (documents XML, données séquentielles, objets...) [Valtchev et al., 2004].

Les liens qui unissent l'ECD et l'AFC, principalement dans l'extraction de motifs fréquents et la génération des règles d'association, seront décrits plus en détail dans la section 2.4.

1.7.3 Représentation des connaissances, Raisonnement

Représentation des connaissances et AFC

La représentation des connaissances est l'une des applications les plus naturelles de l'AFC. En reprenant la définition en cinq points de la représentation des connaissances par Davis *et al.*[Davis et al., 1993], Stumme [Stumme, 2002b] retrace ses liens avec l'AFC.

La représentation des connaissances est un moyen d'expression. Pour Wille [Wille, 1982], le but fondamental de l'AFC est de promouvoir la communication entre les théoriciens et les utilisateurs potentiels de la théorie des treillis. Wille situe cette démarche dans la lignée de l'action communicative d'Habermas [Habermas, 1981]. Ce dernier observe que même la transmission des faits scientifiques procède d'un échange d'arguments et contre-arguments entre plusieurs interlocuteurs. Un des rôles de la représentation de la connaissance est de fournir un support à

cet échange.

La représentation des connaissances est un ensemble de choix ontologiques. En choisissant un modèle de représentation, nous prenons un ensemble de décisions sur ce que doit contenir le monde et la façon dont nous l'appréhendons. Comme nous l'avons décrit dans la section 1.7.1, l'AFC formalise les notions de concept philosophique, extension et intension pour en faire des unités élémentaires de connaissance, sur lesquelles des entités plus complexes peuvent être bâties. En comparant l'AFC à la norme ISO 704 sur les terminologies qui distingue trois niveaux (objet, concept, représentation), Stumme constate que l'AFC se concentre plus sur le niveau concept. Néanmoins, l'AFC peut être un complément de mécanismes plus proches du niveau représentation tels que les logiques de description [Hacene et al., 2007b].

La représentation des connaissances est un substitut. L'homme, ou la machine ne peuvent raisonner directement avec les objets réels. En AFC, ce substitut est incarné par les concepts et treillis de concepts formels.

La représentation des connaissances est une théorie fragmentaire du raisonnement intelligent. Si l'AFC ne propose pas de mécanisme de raisonnement, elle peut servir de base pour alimenter d'autres formalismes comme par exemple les règles d'association [Pasquier et al., 1999c] ou la génération d'hypothèses de la méthode JSM [Ganter and Kuznetsov, 2000].

La représentation des connaissances est un moyen de calcul pragmatique et efficient. Pour cela, un système de représentation des connaissances doit fournir un ensemble de structures qui organise l'information de manière à faciliter ou suggérer le calcul des inférences. Les treillis de concepts et leurs propriétés offrent des mécanismes de calcul efficaces qui ont été particulièrement exploités dans des tâches d'extraction des connaissances [Mineau and Godin, 1995, Pasquier et al., 1999c, Stumme, 2002a]

Ontologies et AFC

En tant qu'unités élémentaires de connaissance, les concepts sont au centre des formalismes de représentation des connaissances, et en particulier des ontologies. Les ontologies sont une spécification explicite, formelle, d'une conceptualisation d'un domaine de connaissance [Gruber, 1995]. Pour Cimiano *et al.* [Cimiano et al., 2004], l'AFC et les ontologies sont deux formalismes complémentaires :

- l'élaboration d'ontologies peut bénéficier de l'AFC comme outil d'apprentissage,
- une ontologie peut être analysée et explorée par des techniques issues de l'AFC,
- des applications basées sur l'AFC peuvent être améliorées par l'utilisation d'ontologies.

Que ce soit dans son esprit, dans ses actions ou productions intellectuelles, la conception d'un domaine par l'homme est souvent implicite. Rendre explicite ce savoir est une des principales préoccupations des concepteurs d'ontologies. Or, pour un domaine donné, ce processus peut

conduire à des résultats différents selon le point de vue des concepteurs, la finalité de l'ontologie, le niveau de granularité... Combiner ces points de vue, c'est-à-dire fusionner des ontologies, demeure un problème d'importance en représentation des connaissances. Aujourd'hui, il n'est pas concevable que cette opération puisse être effectuée sans intervention humaine et la plupart des applications de fusion d'ontologies visent à assister le concepteur dans sa tâche. Stumme et Maedche [Stumme and Maedche, 2001] ont proposé `FCA-MERGE`, un système qui transforme deux ontologies en contextes formels, les fusionne et produit un treillis de concepts qui permet de guider l'utilisateur dans ses choix.

Lorsque la connaissance à formaliser est contenue dans des textes, l'AFC peut être couplée à des méthodes de traitement automatique du langage naturel pour construire une ontologie. Cimiano *et al.* [Cimiano et al., 2005] dérivent un contexte formel à partir d'un corpus en se basant sur les dépendances syntaxiques entre sujets, verbes et compléments. Ils transforment ensuite la relation d'ordre du treillis de concepts associé pour obtenir une hiérarchie conceptuelle. Ce mécanisme est illustré sur la figure 1.13.

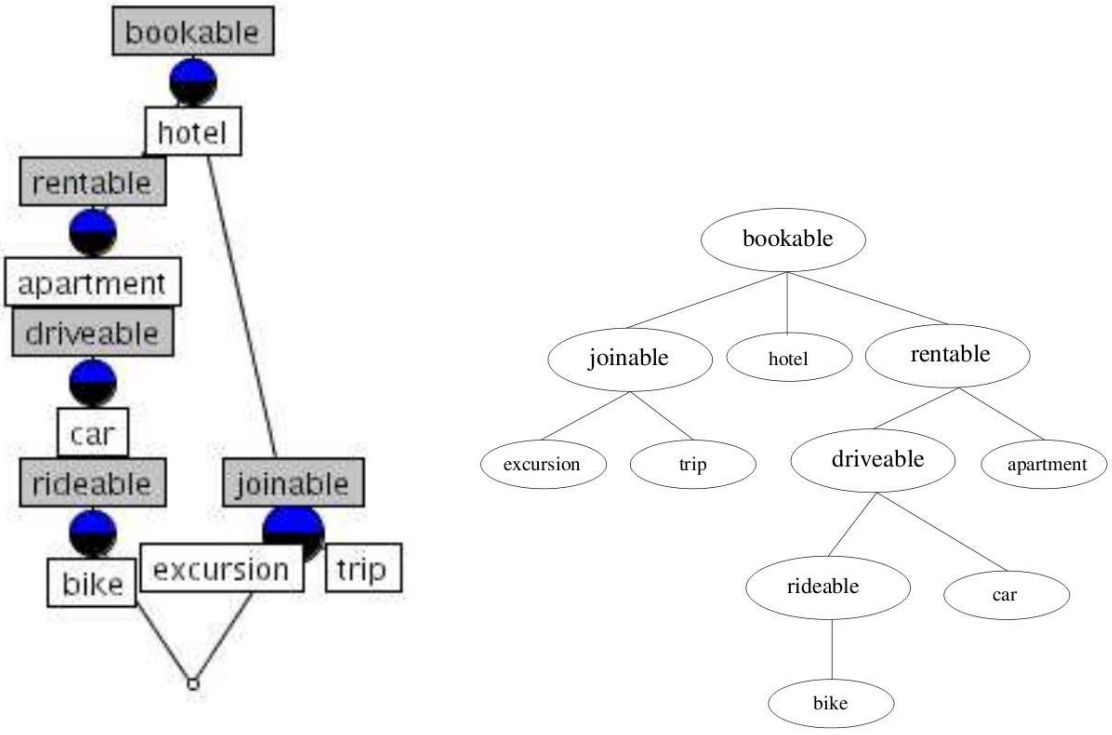

FIG. 1.13 – Construction d'ontologie à partir de textes : transformation d'un treillis de concepts en hiérarchie conceptuelle (d'après [Cimiano et al., 2005]).

Mais, l'AFC peut également donner naissance à des représentations plus expressives. Hacene *et al.* [Hacene et al., 2007b] ont montré les liens existant entre attributs relationnels en ARC et rôles en logiques de description [Baader and Nutt, 2003]. Bendaoud *et al.* ont utilisé l'analyse relationnelle de concepts pour construire automatiquement une ontologie qui puisse être représentée en logiques de description [Bendaoud et al., 2007b].

Les possibilités de navigation offertes par les treillis de concepts, déjà évoquées dans le domaine de la recherche d'information (cf 1.7.4), ont été mises en œuvre sur la plateforme d'apprentissage en E-learning KAON (Karlsruhe Ontology and Semantic Web Framework) [Schmitz et al., 2002]. Ce système utilise des ontologies pour permettre aux utilisateurs d'accéder à des pages Web et des ressources stockées dans un environnement distribué. De manière à faciliter la recherche de documents, la navigation dans ces ontologies est supportée par des treillis de concepts. Pour une ontologie donnée, le système permet de visualiser un treillis construit à partir du contexte formel $(\mathtt{C} \cup \mathtt{I}, \mathtt{C}, \leq_{\mathtt{C}} \cup\ \mathtt{is}-\mathtt{a})$, ou $\mathtt{C}$ est l'ensemble des concepts ontologiques, $\leq_{\mathtt{C}}$ la hiérarchie de subsomption l'ontologie, $\mathtt{I}$ l'ensemble des instances et $\mathtt{is}-\mathtt{a}$ la relation d'instanciation entre $\mathtt{I}$ et $\mathtt{C}$.

1.7.4 Recherche d'information et fouille de texte

D'après [Carpineto and Romano, 2005], l'utilisation de l'AFC en recherche d'information (RI) a comporté trois phases. Les premiers travaux menés par l'équipe de Godin dans les années 80 [Godin et al., 1986, Godin et al., 1989] reposent sur l'idée qu'une requête peut être vue comme l'intension d'un concept formel, le concept-requête. L'extension du concept-requête représente le résultat de la requête, c'est-à-dire l'ensemble des documents recherchés. Le concept-requête est inséré dans un treillis construit à partir de la relation entre une collection de documents et un ensemble de termes d'indexation. Les années 90 ont vu l'intégration de l'AFC avec des techniques classiques de RI pour améliorer les possibilités d'interaction utilisateur-système, voire explorer des méthodes de fouille de texte. Ces dernières années, les chercheurs se sont intéressés aux problèmes de positionnement (classement par pertinence : text ranking) ou encore aux données semi-structurées. De nouveaux domaines ont été explorés comme le courriel, les documents Web et les systèmes de fichiers. Pour Carpineto et Romano, les apports de l'AFC sont multiples tant pour la RI que pour la fouille de texte [Carpineto and Romano, 2005].

Le premier concerne la reformulation des requêtes booléennes. Avec le système REFINER [Carpineto and Romano, 1998], les auteurs construisent une partie du treillis centrée autour du concept correspondant à la requête, qui est le plus général contenant tous les termes de la requête. L'utilisateur peut alors affiner sa recherche en explorant les prédécesseurs ou successeurs directs du concept requête. Des opérations plus complexes permettent de borner l'espace de recherche à l'aide de la relation d'ordre sur les concepts. A partir d'un concept donné

C, on peut, par exemple, élaguer le treillis des résultats en éliminant les concepts inférieurs ou incomparables avec C.

De plus, les propriétés structurelles des treillis de concepts offrent la possibilité de combiner requêtes booléennes et parcours hiérarchique dans une même structure de données, donnant ainsi naissance à des systèmes intégrés de requête et de navigation. L'intégration de thésaurus dans un système d'information garantit également l'indexation des requêtes par les termes les plus généraux et permet de profiter des relations sémantiques apportées par le thésaurus.

L'intégration de connaissances du domaine, sous des formes plus complexes, peut également enrichir la recherche d'information par treillis de concepts : Messai *et al.* [Messai et al., 2007] proposent un système de recherche d'information guidée par les connaissances fondé sur des règles de dépendance entre attributs.

Lorsque les documents contiennent des données structurées, l'AFC permet d'explorer une collection selon de multiples points de vue en utilisant notamment des treillis entrelacés. C'est par exemple la méthode implémentée dans CEM [Cole and Stumme, 2000], un système de gestion de courriels.

Enfin, des travaux se sont intéressés au problème de positionnement des documents qui consiste à classer les résultats d'une requête par ordre de pertinence. L'approche par AFC consiste à mesurer la pertinence des résultats par une métrique basée sur leur distance avec le concept-requête dans le treillis.

1.7.5 Linguistique

Selon Priss [Priss, 2005], on peut distinguer trois domaines d'application de l'AFC en linguistique :

- l'analyse des traits linguistiques (phonétiques, sémantiques, syntaxiques ou grammaticaux),
- la construction automatique ou semi-automatique de bases de données lexicales,
- la représentation et l'analyse de bases de données lexicales.

Les linguistes utilisent fréquemment certains traits pour caractériser un ensemble de termes. Ces traits linguistiques peuvent facilement être interprétés à l'aide l'AFC. Parmi eux, on trouve les composants sémantiques ou sèmes qui sont des unités minimales de signification. Kipke et Wille [Kipke and Wille, 1987] proposent différents treillis de concepts pour modéliser les champs sémantiques couverts par la notion « bodies of water ». L'un d'eux est représenté dans la figure 1.14.

Le traitement automatique du langage naturel repose sur la construction préalable de lexiques qui permettent de dénombrer les mots d'un domaine et de leur associer un ou plusieurs sens. L'élaboration de tels lexiques est une tâche longue et coûteuse. La linguistique

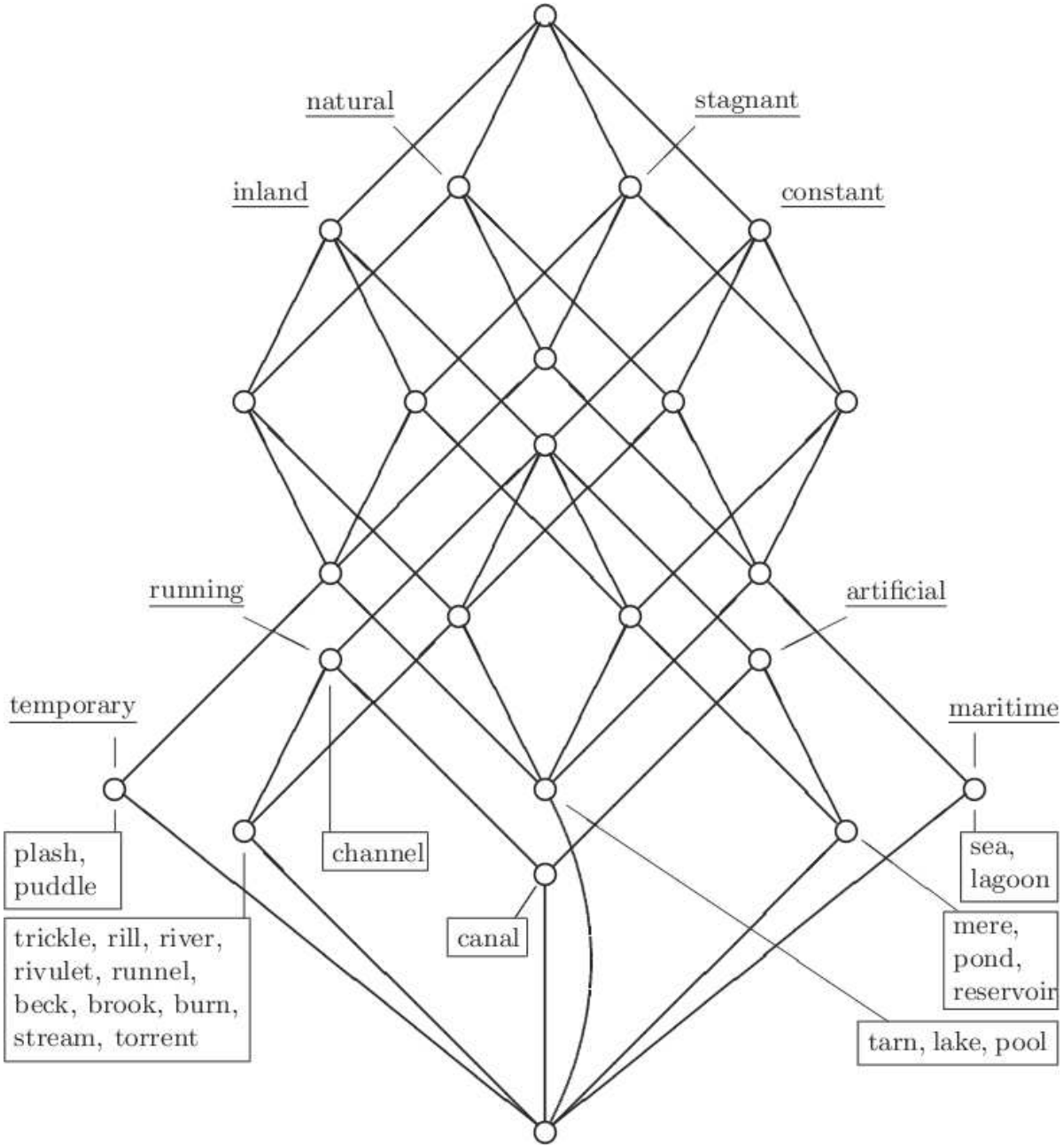

FIG. 1.14 – Treillis des champs sémantiques recouverts par « bodies of water » (d'après [Kipke and Wille, 1987]).

computationelle permet d'accélérer cette étape en s'appuyant notamment sur des corpus. L'un des principaux problèmes qu'elle doit surmonter est de lever l'ambiguïté posée par certains énoncés. Par exemple, dans la phrase « La vitesse moyenne du pigeon voyageur n'est dépassée que par le *vol* de l'hirondelle », le mot *vol* correspond à un déplacement actif dans l'air. Or, *vol* peut aussi signifier : action de dérober. Dans un domaine donné, la construction de lexiques peut s'appuyer sur une ontologie de ce domaine, cette dernière permettant par exemple de découvrir le sens d'un mot à partir de la structure syntaxique de l'énoncé. Mais la linguistique se trouve confrontée à un cercle vicieux car l'élaboration semi-automatique d'ontologies repose elle-même sur l'existence de lexiques détaillés. Pour briser ce cercle, Basili *et al.* [Basili et al., 1997] propose une méthode de raffinement lexical (*lexical tuning*) basée sur l'AFC. Cette méthode consiste à améliorer un lexique par un apprentissage non supervisé à partir d'un corpus. Le lexique obtenu est ensuite réutilisé pour analyser le corpus et ainsi de suite. La phase d'apprentissage repose sur la construction de treillis de concepts qui permettent de désambiguïser des verbes en fonction de leurs arguments dans les phrases qui les contiennent. Selon les auteurs, la méthode peut être utilisée par un expert humain comme outil d'aide à la décision. Alors que l'on estime à une heure le temps nécessaire à la définition d'une entrée lexicale, le gain ainsi obtenu peut être substantiel.

1.7.6 Génie logiciel

Dans [Tilley et al., 2005], les auteurs proposent une revue de la littérature à partir de 47 publications sur l'utilisation de l'AFC dans le domaine du génie logiciel. La figure 1.15 montre un treillis de concepts qui classe ces 47 publications au regard des 13 activités du cycle de vie du développement logiciel, tel qu'il est décrit par la norme ISO 12207.

Le treillis montre que 27 publications décrivent des applications de l'AFC pour la maintenance logicielle (software maintenance) et la conception détaillée (detailed design). La maintenance logicielle est le processus de modification d'un système ou d'un composant après livraison pour la correction de fautes, l'amélioration des performances ou l'adaptation à un nouvel environnement. Typiquement, il s'agit d'utiliser l'AFC pour l'identification de classes dans les systèmes légataires ou la maintenance de hiérarchies de classes dans les systèmes orientés objets contemporains[Snelting and Tip, 1998, Snelting and Tip, 2000, Hacene et al., 2007a, Godin and Mili, 1993].

Le deuxième champ d'application de l'AFC concerne l'analyse des besoins (requirement analysis). Cette phase consiste à étudier les besoins des utilisateurs futurs pour définir les exigences système, matérielles et logicielles. L'AFC peut être utilisée pour la détection de classes au travers de cas d'utilisation [Düwel and Hesse., 1998], ou encore l'extraction de composants pour la réutilisation [Lindig, 1995].

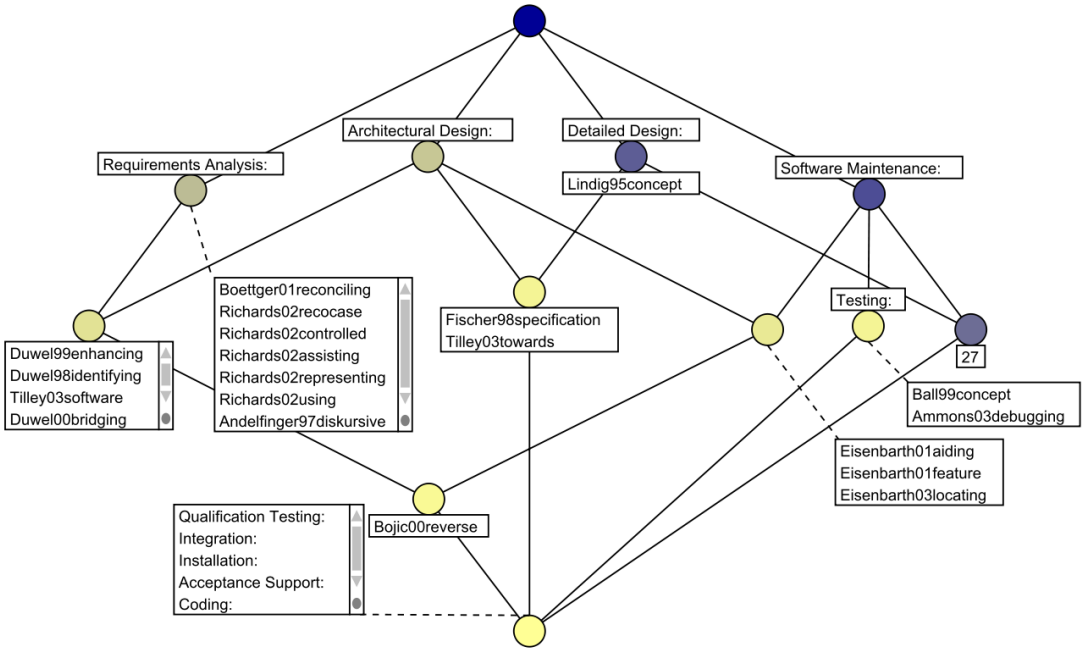

FIG. 1.15 – Treillis de concepts montrant 47 publications traitant de l'AFC dans le domaine du génie logiciel (d'après [Tilley et al., 2005]).

1.8 Conclusion

Initialement introduite comme une restructuration de la théorie des treillis, l'AFC a rapidement dépassé ce cadre. Elle a été utilisée avec succès dans de nombreux domaines : représentation des connaisances, raisonnement, recherche d'information, linguistique, génie logiciel. Plus récemment, elle a fait la preuve de son intérêt dans le champ de l'ECD. Les liens qui l'unissent à la recherche de motifs fréquents et la découverte de règles d'association ont permis d'élaborer des algorithmes de fouille et des méthodes d'interprétation performants que nous présentons dans le chapitre suivant.

Chapitre 2

Extraction de connaissances par extraction de motifs

Sommaire

2.1 Introduction

L'Extraction de Connaissances à partir de Données (ECD) recouvre de très nombreuses méthodes suivant la nature et la complexité des données à traiter. S'agissant de données symboliques, une des plus populaires est la recherche de règles d'association, qui est elle-même une extension du problème de l'extraction de motifs fréquents [Agrawal et al., 1993]. Or, il existe des liens très étroits entre l'AFC et ces deux méthodes. D'une part, les propriétés mathématiques des treillis de concepts sont à la base de plusieurs algorithmes efficients de recherche de motifs fréquents dans de grands volumes de données. D'autre part, l'AFC est à l'origine d'une représentation condensée des règles d'association, et ceci sans perte d'information.

Dans ce chapitre, la section suivante est consacrée à la présentation des motifs fréquents.

Ensuite nous introduisons la recherche de motifs séquentiels fréquents, qui sont une extension des motifs fréquents à des données séquentielles (par exemple ordonnées dans le temps). Enfin, la dernière section établit le lien entre AFC et motifs fréquents. Elle pose par ailleurs la question de la visualisation des motifs par treillis de concepts.

2.2 Recherche de motifs fréquents

Le problème de l'extraction de motifs fréquents dans une base de données a été introduit par Agrawal *et. al* [Agrawal et al., 1993] en 1993. Il consiste en la recherche des ensembles d'éléments présents en même temps dans un nombre suffisamment grand de lignes de la base. Cette méthode permet par exemple d'analyser des données de type « panier de la ménagère ». Chaque ligne de la base représente une *transaction* dans laquelle figure une liste d'articles ou *items*. L'extraction de motifs fréquents de cette base met en évidence les articles fréquemment achetés ensemble.

2.2.1 Cadre formel

Comme nous le verrons dans la section 2.4, le problème d'extraction de motifs fréquents peut être formulé en termes d'AFC. Néanmoins, nous reprenons ici le formalisme habituellement utilisé dans la littérature spécifique de ce domaine.

Définition 23 (Base de données)**.** *Soit* $\mathtt{O}$ *un ensemble fini d'objets,* $\mathtt{I}$ *un ensemble fini d'éléments ou items et* $\mathcal{R}$ *une relation binaire entre* $\mathtt{O}$ *et* $\mathtt{I}$*. On appelle base de données le triplet* $\mathtt{D} = (\mathtt{O}, \mathtt{I}, \mathtt{R})$.

Le tableau 2.1 montre une base de données de quatre objets $(\mathtt{1}, \ldots, \mathtt{4})$ et cinq items $(\mathtt{a}, \ldots, \mathtt{d})$.

Objets	Items				
1	a		c		
2		b	c		
3	a				d
4	a		c		

TAB. 2.1 – Une base de données $\mathtt{D}$.

Définition 24 (Motif)**.** *Soit* $\mathtt{D} = (\mathtt{O}, \mathtt{I}, \mathtt{R})$ *une base de données. On appelle motif tout sous-ensemble de* $\mathtt{I}$*. Un motif* $\mathtt{M}$ *est supporté par un objet* $\mathtt{o}$ *si :*

$$\forall \mathtt{i} \in \mathtt{M}, (\mathtt{o}, \mathtt{i}) \in \mathtt{R}$$

On appelle $|\mathtt{M}|$ *la taille d'un motif. Un motif* $\mathtt{M_1}$ *est un sous-motif (super-motif) d'un motif* $\mathtt{M_2}$ *si* $\mathtt{M_1} \subseteq \mathtt{M_2}$ *(*$\mathtt{M_1} \supseteq \mathtt{M_2}$*).*

Dans la base de données du tableau 2.1, l'objet 1 supporte le motif $\{\mathtt{a}, \mathtt{c}\}$.

Définition 25 (Image d'un motif)**.** *Soit* $\mathtt{D} = (\mathtt{O}, \mathtt{I}, \mathtt{R})$ *une base de données. Soit* $\mathtt{f}$ *la fonction qui fait correspondre à un motif l'ensemble des objets qui le supportent. L'image d'un motif* $\mathtt{M}$ *est :*

$$\mathtt{f}(\mathtt{M}) = \{\mathtt{o} \in \mathtt{O} / \mathtt{o} \text{ supporte } \mathtt{M}\}$$

L'image du motif $\{\mathtt{a}, \mathtt{c}\}$ dans la base de données $\mathtt{D}$ (tableau 2.1) est $\{1, 4\}$.

Définition 26 (Support)**.** *Soit* $\mathtt{D} = (\mathtt{O}, \mathtt{I}, \mathtt{R})$ *une base de données. On appelle support d'un motif* $\mathtt{M}$ *dans* $\mathtt{D}$ *et ont note* $\sigma_{\mathtt{D}}(\mathtt{M})$ *la proportion d'objets de* $\mathtt{O}$ *qui supportent* $\mathtt{M}$ *:*

$$\sigma_{\mathtt{D}}(\mathtt{M}) = \frac{|\mathtt{f}(\mathtt{M})|}{|\mathtt{O}|}$$

On appelle support absolu le cardinal de $\mathtt{f}(\mathtt{M})$*.*

Le support du motif $\{\mathtt{a}, \mathtt{c}\}$ dans $\mathtt{D}$ est $2/4$=0.5.

2.2.2 Le problème de l'extraction des motifs fréquents

Définition 27 (Extraction des motifs fréquents)**.** *Soit* $\mathtt{D} = (\mathtt{O}, \mathtt{I}, \mathtt{R})$ *une base de données, et un seuil ou support minimum* $\mathtt{minsupp} \in [0; 1]$*. Le problème de l'extraction des motifs fréquents consiste à découvrir tous les motifs* $\mathtt{M} \in 2^{\mathtt{I}}$ *tels que :*

$$\sigma_{\mathtt{D}}(\mathtt{M}) \geq \mathtt{minsupp}$$

Ces motifs sont appelés motifs fréquents.

Si on utilise un support minimum de 0.5 pour extraire les motifs fréquents de la base de données de la figure 2.1, on obtient :

- $\{\mathtt{a}\}$ de support $3/4$
- $\{\mathtt{c}\}$ de support $3/4$
- $\{\mathtt{a}, \mathtt{c}\}$ de support $2/4$

D'un point de vue algorithmique, une approche naïve consisterait, pour une base de données $\mathtt{D} = (\mathtt{O}, \mathtt{I}, \mathtt{R})$, à calculer le support de tous les éléments de $2^{\mathtt{I}}$. La complexité de cette approche est rendue exponentielle par l'explosion combinatoire du nombre de motifs candidats. Les algorithmes d'extraction de motifs fréquents adoptent principalement deux stratégies pour résoudre ce problème efficacement dans de grands volumes de données :

- la réduction du nombre de motifs candidats,
- la réduction de l'espace de recherche utilisé pour le calcul du support. Elle permet de réduire les temps d'accès à la base de données, surtout lorsque celle-ci est stockée sur un disque.

Une propriété fondamentale permettant de réduire le nombre de motifs candidats est la suivante :

Propriété 8. *Pour qu'un motif soit fréquent, tous ses sous-motifs doivent être fréquents.*

`Apriori` [Agrawal and Srikant, 1994] est l'un des tout premiers algorithmes de recherche de motifs fréquents. Son principe est de trouver les motifs fréquents de taille `k` en combinant les motifs fréquents de taille $\mathtt{k}-\mathtt{1}$. `Apriori` tire ainsi parti de la propriété 8 pour élaguer l'espace de recherche. L'extraction de motifs fréquents a fait l'objet de recherches très actives depuis une quinzaine d'années, et de nombreux algorithmes ont été proposés [Han et al., 2007]. Il existe des connections étroites entre l'AFC et l'extraction de motifs fréquents, qui ont donné naissance à plusieurs algorithmes et que nous détaillons dans la section 2.4.

2.3 Motifs séquentiels fréquents

2.3.1 Introduction

Introduite par Agrawal et Srikant [Agrawal and Srikant, 1995], l'extraction de motifs séquentiels fréquents est une méthode de fouille de données destinée à découvrir de façon automatique des répétitions caractéristiques dans une base de données ordonnées. Ce peut être par exemple une succession de bases ATGC que l'on retrouve fréquemment dans le même ordre sur différents brins d'ADN, ou encore un comportement identique de lecteurs dans une bibliothèque qui emprunteraient successivement et dans le même ordre *Les Trois Mousquetaires* puis *Vingt ans après* et enfin *Le Vicomte de Bragelonne* de la trilogie de Dumas.

2.3.2 Cadre formel

Définition 28 (Séquence)**.** *Soit* `I` *un ensemble fini d'items. Une séquence engendrée par* `I` *est un liste ordonnée de parties de* `I`*.*

Notations :

- $\mathtt{S} = \langle \mathtt{s_1 s_2 \dots s_m} \rangle$ est une *séquence* dans laquelle les $\mathtt{s_i}$ sont des *parties* de `I`. On dit également que les $\mathtt{s_i}$ sont des parties de `S`.
- On note $(\mathtt{i_1 i_2 \dots i_n})$ une *partie* de séquence dans laquelle les $\mathtt{i_j}$ sont des *éléments* de `I`.

- On appelle taille d'une séquence $\mathtt{S}$ et on note $|\mathtt{S}|$ le nombre d'items dans $\mathtt{S}$. Une séquence de taille k est une k-séquence.
- On appelle longueur d'une séquence le nombre de parties de cette séquence.

Soit $\mathtt{I} = \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}, \mathtt{e}\}$ un ensemble d'items.

- $\mathtt{S} = \langle(\mathtt{a})(\mathtt{ae})(\mathtt{cde})\rangle$ est une séquence
- $(\mathtt{ae})$ est une partie de $\mathtt{S}$
- $|\mathtt{S}| = 1 + 2 + 3 = 6$.
- la longueur de $|\mathtt{S}|$ est 3.

Définition 29 (Sous-séquence). *Une séquence* $\mathtt{A} = \langle \mathtt{a}_1\mathtt{a}_2 \ldots \mathtt{a}_n\rangle$ *est une sous-séquence de* $\mathtt{B} = \langle \mathtt{b}_1\mathtt{b}_2 \ldots \mathtt{b}_m\rangle$ *s'il existe n entiers* $\mathtt{i}_1, \mathtt{i}_2, \ldots \mathtt{i}_n$ *tels que :*

- $\mathtt{i}_1 < \mathtt{i}_2 < \ldots < \mathtt{i}_n$ *et*
- $\mathtt{a}_1 \subset \mathtt{b}_{\mathtt{i}_1}, \mathtt{a}_2 \subset \mathtt{b}_{\mathtt{i}_2} \ldots \mathtt{a}_n \subset \mathtt{b}_{\mathtt{i}_n}$

On note $\mathtt{A} \subseteq \mathtt{B}$. *On dit également que* $\mathtt{B}$ *supporte* $\mathtt{A}$.

Soit la séquence $\mathtt{S} = \langle(\mathtt{c})(\mathtt{abc})(\mathtt{bd})\rangle$.

- $\langle(\mathtt{ab})(\mathtt{bd})\rangle$ est une sous-séquence de $\mathtt{S}$,
- $\langle(\mathtt{c})(\mathtt{ad})\rangle$ n'est pas une sous-séquence de $\mathtt{S}$.

Définition 30 (Super-séquence). *Une séquence* $\mathtt{S}$ *est une super-séquence de* $\mathtt{S}'$ *si et seulement si* $\mathtt{S}'$ *est une sous-séquence de* $\mathtt{S}$.

Définition 31 (Base de séquences). *Soient* $\mathtt{O}$ *un ensemble fini d'objets,* $\mathtt{I}$ *un ensemble fini d'items et* Σ *l'ensemble des séquences engendrées par* $\mathtt{I}$. *Soit* $\mathtt{R}$ *une relation binaire entre* $\mathtt{O}$ *et* Σ *telle que* $\forall\, \mathtt{o} \in \mathtt{O}, \exists!\mathtt{S} \in \Sigma |\, (\mathtt{o}, \mathtt{S}) \in \mathtt{R}$. *On appelle base de séquences le quadruplet* $\mathtt{D} = (\mathtt{O}, \mathtt{I}, \Sigma, \mathtt{R})$.

Soit $\mathtt{O} = \{\mathtt{p}_1, \mathtt{p}_2, \mathtt{p}_3, \mathtt{p}_4\}$ un ensemble de quatre patients. On reconstitue pour chacun la séquence de traitements qu'il a reçus parmi l'ensemble des traitements $\mathtt{I} = \{\mathtt{x}, \mathtt{r}, \mathtt{c}\}$ où $\mathtt{x}$ est la chirurgie, $\mathtt{r}$ une cure de radiothérapie, $\mathtt{c}$ une cure de chimiothérapie. Les séquences de traitements sont les suivantes :

- $\mathtt{p}_1$ a reçu $\mathtt{x}$ puis $\mathtt{c}$ puis $\mathtt{c}$ et à nouveau $\mathtt{x}$,
- $\mathtt{p}_2$ a reçu $\mathtt{x}$ et $\mathtt{r}$ de façon concomitante puis $\mathtt{c}$,
- $\mathtt{p}_3$ a reçu $\mathtt{c}$ puis $\mathtt{c}$ puis $\mathtt{x}$,
- $\mathtt{p}_4$ a reçu $\mathtt{c}$ puis $\mathtt{c}$ puis $\mathtt{r}$ puis $\mathtt{r}$.

La base de séquences $\mathtt{D}$ des traitements des patients de $\mathtt{O}$ est représentée dans le tableau 2.2.

Définition 32 (Image d'une séquence). *Soit* $\mathtt{f}$ *la fonction qui associe à une séquence* $\mathtt{S}$ *de* Σ *l'ensemble des objets* $\mathtt{o}$ *de* $\mathtt{O}$ *qui sont associés à une super-séquence de* $\mathtt{S}$

$$\begin{aligned} \mathtt{f} : \Sigma &\rightarrow \mathtt{O} \\ \mathtt{S} &\mapsto \{\mathtt{o} / \exists\, \mathtt{S}' \in \Sigma, (\mathtt{o}, \mathtt{S}') \in \mathtt{R}, \mathtt{S} \subseteq \mathtt{S}'\} \end{aligned}$$

patients	séquences
$\mathtt{p}_1$	$\langle(\mathtt{x})(\mathtt{c})(\mathtt{c})(\mathtt{x})\rangle$
$\mathtt{p}_2$	$\langle(\mathtt{xr})(\mathtt{c})\rangle$
$\mathtt{p}_3$	$\langle(\mathtt{c})(\mathtt{c})(\mathtt{x})\rangle$
$\mathtt{p}_1$	$\langle(\mathtt{c})(\mathtt{c})(\mathtt{r})(\mathtt{r})\rangle$

TAB. 2.2 – Base $\mathtt{D}$ des séquences de traitements reçus par les patients de l'ensemble $\mathtt{O}$.

On dit qu'un objet $\mathtt{o}$ *supporte une séquence* $\mathtt{S}$ *si* $\mathtt{o} \in \mathtt{f}(\mathtt{S})$. *La séquence vide* $\emptyset = \langle\rangle$ *est supportée par tous les objets de* $\mathtt{O}$.

Soit la séquence $\mathtt{S} = \langle(\mathtt{x})(\mathtt{c})\rangle$. Dans la base de séquences du tableau 2.2, $\mathtt{f}(\mathtt{S}) = \{\mathtt{p}_1, \mathtt{p}_2\}$ représente tous les patients qui ont bénéficié d'une chirurgie, puis d'une chimiothérapie.

Définition 33 (Support d'une séquence)**.** *On appelle support d'une séquence* $\mathtt{S}$ *dans* $\mathtt{D}$ *et on note* $\sigma_{\mathtt{D}}(\mathtt{S})$ *la proportion d'objets de* $\mathtt{O}$ *qui supportent* $\mathtt{S}$.

$$\sigma_{\mathtt{D}}(\mathtt{S}) = \frac{|\mathtt{f}(\mathtt{S})|}{|\mathtt{O}|}$$

$\mathtt{S} = \langle(\mathtt{x})(\mathtt{c})\rangle$ est supportée par deux des quatre patients de $\mathtt{O}$. $\sigma_{\mathtt{D}}(\mathtt{S}) = \frac{2}{4} = 0.5$

Définition 34 (Motif séquentiel fréquent)**.** *Étant donné un seuil* $\mathtt{minsupp} \in [0, 1]$ *et une base de séquences* $\mathtt{D} = (\mathtt{O}, \mathtt{I}, \Sigma, \mathtt{R})$*, une séquence* $\mathtt{S}$ *est un motif séquentiel fréquent si :*

$$\sigma_{\mathtt{D}}(\mathtt{S}) \geq \mathtt{minsupp}$$

Définition 35 (Extraction de motifs séquentiels fréquents)**.** *Le processus d'extraction de motifs séquentiels fréquents dans une base de séquences* $\mathtt{D} = (\mathtt{O}, \mathtt{I}, \Sigma, \mathtt{R})$ *consiste à trouver toutes les séquences de* $\mathtt{S}$ *vérifiant la propriété précédente.*

Si l'on fouille la base $\mathtt{D}$ du tableau 2.2 avec un support minimum de 0.5 on obtient les motifs séquentiels fréquents suivants :

- $\langle(\mathtt{x})\rangle$
- $\langle(\mathtt{r})\rangle$
- $\langle(\mathtt{c})\rangle$
- $\langle(\mathtt{x})(\mathtt{c})\rangle$
- $\langle(\mathtt{c})(\mathtt{x})\rangle$
- $\langle(\mathtt{c})(\mathtt{c})\rangle$
- $\langle(\mathtt{c})(\mathtt{c})(\mathtt{x})\rangle$

2.4 Liens avec l'AFC

Le problème d'extraction des motifs fréquents peut être facilement reformulé en termes d'AFC, la définition d'une base de données étant identique à celle d'un contexte formel. Nous adopterons par conséquent dans la suite de ce document les notations de l'AFC. En particulier, une base de données $\mathtt{D} = (\mathtt{O}, \mathtt{I}, \mathtt{R})$ est assimilée à un contexte formel $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$. La correspondance entre les deux formalismes est donnée dans le tableau 2.3.

	Motifs fréquents	AFC
base de données	$\mathtt{D} = (\mathtt{O}, \mathtt{I}, \mathtt{R})$	$\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$
motif	$\mathtt{M} \subseteq \mathtt{I}$	$\mathtt{Y} \subseteq \mathtt{M}$
image d'un motif	$\mathtt{f}(\mathtt{M})$	$\mathtt{Y}'$
support d'un motif	$\sigma_{\mathtt{D}}(\mathtt{M})$	$\sigma_{\mathbb{K}}(\mathtt{Y}) = \frac{\lvert\mathtt{Y}'\rvert}{\lvert\mathtt{G}\rvert}$

TAB. 2.3 – Correspondances entre formalismes des motifs fréquents et de l'AFC.

Les apports de l'AFC pour la recherche de motifs fréquents et l'extraction de règles d'association sont de deux ordres :

- l'AFC a donné naissance à des algorithmes d'extraction de *motifs fréquents fermés* qui permettent de réduire l'espace de recherche des motifs fréquents,
- l'AFC peut produire des *bases de règles informatives*, qui diminuent le nombre de règles d'associations extraites, sans perte d'information.

2.4.1 Motifs fermés fréquents

En dehors des nombreuses variations de l'algorithme `Apriori`, d'autres approches se sont focalisées sur l'extraction des motifs fréquents maximaux. Un motif fréquent est dit maximal s'il est lui même fréquent et si tous ses super-motifs ne sont pas fréquents. L'ensemble des motifs fréquents peut être simplement dérivé de l'ensemble des motifs fréquents maximaux. L'efficacité des algorithmes qui adoptent cette approche, comme `Max-Miner` [Bayardo, 1998], est cependant limitée par l'étape coûteuse de calcul du support de tous les motifs fréquents après leur génération à partir des motifs fréquents maximaux. Or, des travaux ont montré que l'ensemble des motifs fréquents, ainsi que leur support, peuvent être déduits de l'ensemble des motifs fermés fréquents (MFF) [Pasquier et al., 1999c, Bastide et al., 2000, Stumme et al., 2001, Stumme et al., 2002, Zaki and Hsiao, 2002].

Définition 36 (Motif fermé fréquent)**.** *Soient un contexte formel* $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$*,* $\underline{\mathfrak{B}}(\mathbb{K})$ *son treillis de concepts et* $\mathtt{minsupp} \in [0; 1]$ *un seuil de support minimum. Soit* $\mathtt{Y} \subseteq \mathtt{M}$ *un motif.*

Un concept formel $(\mathtt{A}, \mathtt{B}) \in \underline{\mathfrak{B}}(\mathbb{K})$ *est dit fréquent si :*

$$\sigma_{\mathbb{K}}(\mathtt{B}) \geq \mathtt{minsupp}$$

$\mathtt{Y}$ *est un motif fermé fréquent si :*

- *il est fermé :* $\mathtt{Y}'' = \mathtt{Y}$,
- *et le concept* $(\mathtt{Y}', \mathtt{Y})$ *est fréquent.*

Les MFF possèdent des propriétés très intéressantes pour l'élaboration d'algorithmes de recherche de motifs fréquents. Ils permettent d'abord de réduire l'espace de recherche :

Propriété 9. *Soit* $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$ *un contexte formel et* $\mathtt{B} \subseteq \mathtt{M}$. $\mathtt{B}$ *est fermé si*

$$\forall \mathtt{m} \in \mathtt{M} - \mathtt{B}, \sigma_{\mathbb{K}}(\mathtt{B} \cup \{\mathtt{m}\}) < \sigma_{\mathbb{K}}(\mathtt{B})$$

Autrement dit, un motif est fermé s'il ne peut être augmenté sans réduire son support. D'autre part, les MFF sont une représentation compacte et sans perte d'information de l'ensemble des motifs fréquents, ce qui vient de la propriété suivante :

Propriété 10. *Tout motif possède le même support que sa fermeture.*

Plusieurs algorithmes se basent sur l'extraction des MFF. On peut citer les principaux :

- `Close` [Pasquier et al., 1999c],
- `Apriori-Close` [Pasquier et al., 1999a],
- `A-Close`[Pasquier et al., 1999b],
- `CHARM` [Zaki and Hsiao, 2002],
- `CHARM-L` [Zaki and Hsiao, 2005],
- `Closet`[Pei et al., 2000],
- `Closet+` [Wang et al., 2003],
- `MaxClosure` [Cristofor et al., 2000],
- `TITANIC` [Stumme et al., 2002],
- `PASCAL` [Bastide et al., 2000].

Dans la section suivante, nous nous intéressons à l'algorithme `TITANIC` et à la construction d'icebergs de concepts.

2.4.2 Icebergs de concepts

Les capacités de visualisation de l'AFC sont souvent mises en défaut en présence de contextes formels de grande taille. En introduisant `TITANIC` [Stumme et al., 2002], Stumme *et al.* proposent de dépasser en partie cette limite en ne visualisant que les concepts les plus fréquents dans un iceberg de concepts.

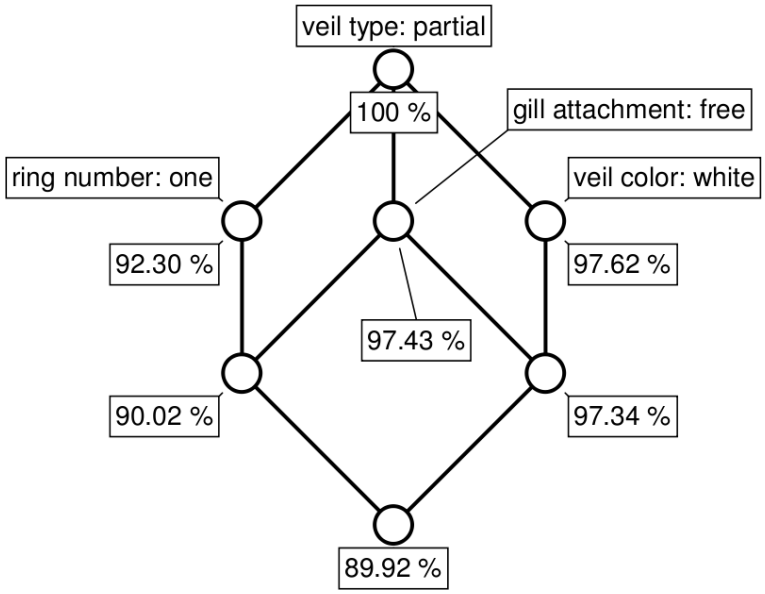

FIG. 2.1 – Iceberg au seuil 85% du contexte MUSHROOMS (d'après [Stumme, 2002a]).

Définition 37 (Iceberg de concepts)**.** *Soit* $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$ *un contexte formel et* `minsupp` *un seuil minimum de support. L'ensemble des concepts fréquents dérivés de* $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$ *et* `minsupp` *est appelé iceberg de concepts.*

Propriété 11. *Un iceberg de concepts a une structure de demi-treillis supérieur.*

Cette propriété découle de la monotonie de la fonction de support : soient $\mathtt{M_1} \subseteq \mathtt{M_2}$ deux ensembles d'attributs d'un contexte $\mathbb{K}$, on a $\sigma_{\mathbb{K}}(\mathtt{M_1}) \geq \sigma_{\mathbb{K}}(\mathtt{M_2})$ De plus, en ajoutant un concept $\perp = (\emptyset, \emptyset)$ à un iceberg, on obtient à nouveau un treillis.

Pour illustrer l'intérêt des icebergs en termes de découverte de connaissances et de visualisation, nous reprenons l'exemple utilisé par Stumme *et al.* sur la base MUSHROOMS [1] Cette base contient 8 416 objets et 22 attributs nominaux équivalents à 80 attributs binaires. Au seuil de 85%, on trouve 16 motifs fréquents dont seulement 7 sont fermés. La figure 2.1 montre l'iceberg de concepts correspondant à ce seuil. Seuls 4 attributs sont fréquents : `veil : partial`, `ringnumber : one`, `gillattachment : free` et `veilcolor : white`. Toutes les combinaisons possibles de ces attributs sont elles-mêmes fréquentes et témoignent de leur forte corrélation. L'iceberg permet de simplifier, hiérarchiser et visualiser la représentation de ces combinaisons. De plus, il peut servir de point de départ au calcul des règles d'associations qui sont une extension naturelle de la recherche de motifs fréquents [Agrawal et al., 1993].

Au delà de la visualisation, `TITANIC` s'appuie sur les propriétés de l'AFC pour une recherche

[1] UCI Machine Learning Repository. http ://www.ics.uci.edu/ learn/MLRepository.html

efficace des motifs fréquents. Comme `Apriori`, `TITANIC` effectue une recherche incrémentale de motifs fréquents. À l'étape `i`, il génère et teste le support de motifs candidats de taille `i` par combinaison des motifs fréquents de taille $\mathtt{i}-\mathtt{1}$ obtenus à l'étape $\mathtt{i}-\mathtt{1}$.

À la différence d'`Apriori`, `TITANIC` ne recherche que les générateurs minimaux fréquents. Il utilise deux propriétés principales. La première est celle qui lie un générateur minimal à sa classe d'équivalence. En connaissant le support d'un générateur minimal, on peut déduire le support de l'ensemble des motifs de sa classe d'équivalence en calculant sa fermeture. Cette propriété permet de limiter les accès à la base de données pour le calcul du support. La seconde propriété est la suivante :

Propriété 12. *L'ensemble des générateurs minimaux d'un contexte* $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$ *est un idéal d'ordre de* $2^{\mathtt{M}}$ *:*

- *tout sous-ensemble d'un générateur minimal est lui même un générateur minimal,*
- *tout super-ensemble d'un ensemble non générateur minimal n'est pas un générateur minimal.*

Cette propriété permet d'élaguer l'espace de recherche en écartant les candidats qui ne sont pas des générateurs minimaux. Pour plus de détail, le lecteur peut se reporter à [Stumme et al., 2002].

2.4.3 Motifs séquentiels fermés fréquents

Le principe de fermeture utilisé dans la recherche de motifs fréquents a été étendu au domaine des motifs séquentiels fréquents. Avec l'algorithme `CloSpan`, Yan *et al.* [Yan et al., 2003] ont introduit la notion de motif séquentiel fermé.

Définition 38 (Motif séquentiel fermé)**.** *Soit* $\mathtt{D} = (\mathtt{O}, \mathtt{I}, \Sigma, \mathtt{R})$ *une base de séquences. Soit* $\mathtt{S}$ *une séquence de* Σ*.* $\mathtt{S}$ *est un motif séquentiel fermé si et seulement si il n'existe pas de super-séquence de* $\mathtt{S}'$ *de même support que* $\mathtt{S}$*.*

Cette définition de la fermeture ne doit pas être confondue à celle qui est utilisée en AFC. En particulier deux motifs séquentiels fermés incomparables par la relation de sous/super séquence peuvent être supportés par le même ensemble d'objets. Considérons la base composée des deux séquences suivantes : $(\mathtt{o}_1, \langle(\mathtt{a})(\mathtt{b})(\mathtt{c})\rangle)$ et $(\mathtt{o}_2, \langle(\mathtt{ab})(\mathtt{c})\rangle)$. Les motifs séquentiels $\langle(\mathtt{a})(\mathtt{c})\rangle$ et $\langle(\mathtt{b})(\mathtt{c})\rangle$ sont fermés au sens de la définition séquentielle : aucune de leurs super-séquences n'a le même support. Bien qu'incomparables, ils sont tous deux supportés par les objets $\mathtt{o}_1$ et $\mathtt{o}_2$.

Cependant, la recherche des Motifs Séquentiels Fermés Fréquents (MSFF) présente les mêmes avantages que celle des MFF :

- compacité : les MSFF sont une représentation condensée, sans perte d'information, des motifs séquentiels fréquents.
- efficacité algorithmique : la propriété de fermeture (au sens séquentiel) permet d'élaguer l'espace de recherche.

2.5 Conclusion

Bien que l'apparition du problème d'extraction des motifs fréquents soit postérieure à celle de l'AFC, les liens qui les unissent n'ont pas été établis immédiatement. Pourtant l'AFC est à la base d'une famille d'algorithmes performants d'extraction de motifs fréquents. Elle permet de réduire l'espace de recherche dans la génération de motifs et de simplifier les résultats sans perte d'information. D'autre part, ses capacités de visualisation en font une méthode élégante d'interprétation des motifs.

Si l'extension de la recherche de motifs fréquents à des données ordonnées utilise également la notion de fermeture, le lien entre AFC et motifs séquentiels n'est pas aussi évident. Nous explorerons dans le chapitre 5 le potentiel de L'AFC pour l'analyse des motifs fréquents séquentiels.

L'extraction et la représentation des connaissances par icebergs de concepts permettent de réduire la complexité de treillis volumineux. Elles reposent sur une sélection d'unités de connaissance potentiellement intéressantes guidée par une mesure de qualité des concepts : le support. Ce principe présume que l'intérêt d'une connaissance nouvelle dépend de sa fréquence. Nous verrons dans le chapitre suivant que cette approche peut être mise en défaut et nous étudierons une autre mesure de qualité des concepts : la stabilité.

Chapitre 3

Mesurer l'intérêt des concepts

Sommaire

3.1 Introduction

Une des limites de l'ECD provient du volume des résultats à analyser après l'étape de fouille. Dans la recherche de motifs fréquents et la découverte de règles d'association, une étape de post-traitement est souvent nécessaire pour dégager les motifs les plus intéressants et les plus pertinents. Ce processus a fait l'objet de nombreux travaux [Guillet and Hamilton, 2007, Cherfi, 2004] qui ont donné naissance à de multiples mesures d'intérêt. Une mesure d'intérêt permet d'ordonner les règles selon un critère et de présenter, en premier lieu, les règles les plus pertinentes, c'est-à-dire des règles qui, potentiellement, présenteraient un intérêt pour l'analyste.

Si dans le domaine des règles d'associations, ces mesures d'intérêt font régulièrement débat, en revanche, en AFC, seules deux mesures ont été proposées pour étudier l'intérêt des concepts : le support et la stabilité. En outre, le support d'un concept est généralement utilisé pour la construction de règles d'associations et non pas pour l'intérêt du concept en tant que tel.

Le support d'un concept a été étudié dans le chapitre précédent. Ici, nous présentons la notion de stabilité d'un concept et proposons une analyse qualitative de cette mesure d'intérêt.

3.2 Stabilité

La notion de stabilité d'un concept formel a été probablement introduite pour la première fois en 1990 par Kuznetsov [Kuznetsov, 1990]. Depuis, d'autres définitions en ont été données par le même auteur [Kuznetsov, 2007, Kuznetsov et al., 2007]. Par la suite, nous nous référerons à la définition suivante [Kuznetsov et al., 2007].

Définition 39. *Soient* $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$ *un contexte formel,* $\underline{\mathfrak{B}}(\mathbb{K})$ *son treillis de concepts et* $\mathtt{C} = (\mathtt{A}, \mathtt{A}')$ *un concept de* $\underline{\mathfrak{B}}(\mathbb{K})$*. On appelle stabilité de* $\mathtt{C}$ *:*

$$\gamma(\mathtt{C}) = \frac{|\{\mathtt{X} \subseteq \mathtt{A} | \mathtt{X}' = \mathtt{A}'\}|}{2^{|\mathtt{A}|}} \tag{3.1}$$

La stabilité d'un concept indique dans quelle mesure l'intension de ce concept est liée à la présence d'objets particuliers dans son extension. Un concept stable possède une existence d'autant plus « réelle » que son intension ne dérive pas fortuitement d'une description bruitée de ses objets. La stabilité traduit la probabilité qu'un concept, et surtout son intension, persiste même si on retire des objets de son extension. Une formulation différente de la stabilité dans la proposition suivante rend cette idée plus intuitive.

Propriété 13 ([Kuznetsov et al., 2007])**.** *Soient* $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$ *un contexte formel,* $\mathfrak{B}(\mathbb{K})$ *son treillis de concepts et* $\mathtt{C} = (\mathtt{B}', \mathtt{B})$ *un concept de* $\mathfrak{B}(\mathbb{K})$*. Soient* $\mathtt{H} \subseteq \mathtt{G}$*,* $\mathtt{I}_\mathtt{H} = \mathtt{I} \cap (\mathtt{H} \times \mathtt{M})$ *et* $\mathbb{K}_\mathtt{H} = (\mathtt{H}, \mathtt{M}, \mathtt{I}_\mathtt{H})$*. On note* $^{\mathtt{I}_\mathtt{H}}$ *l'opérateur de dérivation redéfini pour le contexte formel* $\mathbb{K}_\mathtt{H}$*. On a alors :*

$$\gamma(\mathtt{C}) = \frac{|\{\mathbb{K}_\mathtt{H} | \mathtt{H} \subseteq \mathtt{G} \text{ et } \mathtt{B} = \mathtt{B}^{\mathtt{I}_\mathtt{H}\mathtt{I}_\mathtt{H}}\}|}{2^{|\mathtt{G}|}}$$

Si on considère $\mathtt{H}$ comme une variable aléatoire à valeur dans $2^\mathtt{G}$ suivant une loi uniforme, la stabilité d'un concept $\mathtt{C} = (\mathtt{B}', \mathtt{B})$ est égale à la probabilité $\mathtt{P}(\mathtt{B} = \mathtt{B}^{\mathtt{I}_\mathtt{H}\mathtt{I}_\mathtt{H}})$. Autrement dit, c'est la probabilité de retrouver $\mathtt{B}$ comme intension d'un concept en supprimant aléatoirement des objets de $\mathbb{K}$.

En poussant encore l'analogie avec les probabilités, on peut faire le lien avec les méthodes d'estimation par ré-échantillonnage, et en particulier celle du *jackknife*. Cette méthode consiste à estimer un paramètre $\mathtt{p}$ (par exemple une moyenne, une variance, une médiane...) d'une population en partant d'un échantillon de taille $\mathtt{n}$ de cette population. Le paramètre $\mathtt{p}$ est alors ré-estimé sur des sous-échantillons de taille $\mathtt{n} - 1$ de l'échantillon initial. C'est la technique du *leave-one-out* qui écarte à chaque étape une observation du jeu de données initiales. L'intérêt du jackknife réside dans sa robustesse vis-à-vis des valeurs extrêmes, celles-ci pouvant être assimilées à des observations bruitées ou des exceptions. En quelque sorte, la stabilité d'un

concept mesure la probabilité de son existence réelle en analysant sa persistance sur tous les sous échantillons d'objets du contexte formel initial. Un concept stable a tendance à persister même en présence d'exceptions. En cela, la stabilité permet de mesurer la robustesse d'un concept en présence de valeurs extrêmes ou aberrantes.

Il est également possible de reformuler la stabilité en termes de classe d'équivalence.

Propriété 14. *Soient* $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$ *un contexte formel,* $\underline{\mathfrak{B}}(\mathbb{K})$ *son treillis de concepts et* $\mathtt{C} = (\mathtt{A}, \mathtt{A}')$ *un concept de* $\underline{\mathfrak{B}}(\mathbb{K})$.

$$\gamma(\mathtt{C}) = \frac{|\langle \mathtt{A} \rangle|}{2^{|\mathtt{A}|}} \tag{3.2}$$

La figure 3.1 montre un exemple de calcul de la stabilité dans un treillis de concepts. Chaque concept est étiqueté par sa stabilité, son extension et son intension complètes. Ainsi, pour le concept $(\{\mathtt{1},\mathtt{5},\mathtt{6}\},\{\mathtt{a}\})$, on a :

$$\begin{aligned}
\emptyset' = \{\mathtt{a},\mathtt{b},\mathtt{c},\mathtt{d}\} &\neq \{\mathtt{a}\} \\
\{\mathtt{1}\}' &= \{\mathtt{a}\} \\
\{\mathtt{5}\}' &= \{\mathtt{a}\} \\
\{\mathtt{6}\}' = \{\mathtt{a},\mathtt{b},\mathtt{c}\} &\neq \{\mathtt{a}\} \\
\{\mathtt{1},\mathtt{5}\}' &= \{\mathtt{a}\} \\
\{\mathtt{1},\mathtt{6}\}' &= \{\mathtt{a}\} \\
\{\mathtt{5},\mathtt{6}\}' &= \{\mathtt{a}\} \\
\{\mathtt{1},\mathtt{5},\mathtt{6}\}' &= \{\mathtt{a}\}
\end{aligned}$$

Ce qui donne $\gamma(\{\mathtt{1},\mathtt{5},\mathtt{6}\},\{\mathtt{a}\}) = \frac{\mathtt{6}}{\mathtt{8}} = \mathtt{0.75}$. On peut noter que la stabilité est (par définition) toujours comprise entre 0 et 1. Notons par ailleurs que $\gamma(\bot) = \mathtt{1}$. Cette égalité est toujours vérifiée puisque tout sous-ensemble de l'extension de $\bot$ a pour image l'intension de $\bot$. De plus, la stabilité d'un concept est toujours strictement positive, puisque son extension contient au moins $\emptyset$.

La stabilité d'un concept peut être vue comme une mesure qui s'applique à l'idéal de ce concept dans le treillis, comme le montre la propriété suivante.

Propriété 15. *Soient* $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$ *un contexte formel,* $\underline{\mathfrak{B}}(\mathbb{K})$ *son treillis de concepts, et* $\mathfrak{U}(\mathbb{K})$ *l'ensemble de leurs extensions. Soit* $\mathtt{C} = (\mathtt{A}, \mathtt{A}')$ *un concept de* $\underline{\mathfrak{B}}(\mathbb{K})$. *Si on note* $\mathfrak{U}_{\mathtt{A}}$ *l'ensemble*

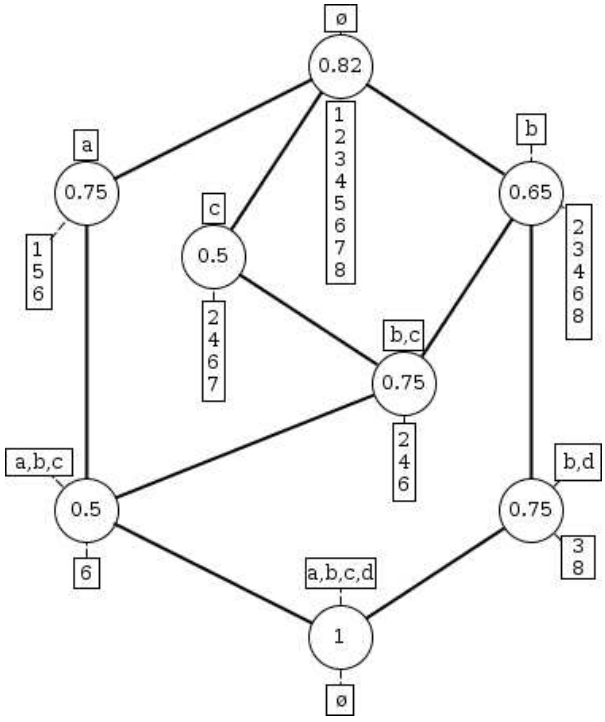

FIG. 3.1 – Exemple de calcul de la stabilité dans un treillis

$(\mathfrak{U} \cap 2^{\mathtt{A}}) - \{\mathtt{A}\}$, *alors on a :*

$$\gamma(\mathtt{C}) = 1 - \sum_{\mathtt{X} \in \mathfrak{U}_{\mathtt{A}}} \gamma(\mathtt{X}, \mathtt{X}') 2^{|\mathtt{X}| - |\mathtt{A}|} \tag{3.3}$$

Preuve. Pour un concept $\mathtt{C} = (\mathtt{A}, \mathtt{A}')$, la propriété (14), donne :

$$\gamma(\mathtt{C}) = \frac{|\langle A \rangle|}{2^{|A|}}$$

$\mathfrak{U}_{\mathtt{A}} \cup \{\mathtt{A}\}$ est l'ensemble des extensions des concepts de l'idéal de $\mathtt{C}$ dans $\underline{\mathfrak{B}}(\mathbb{K})$. L'ensemble des classes d'équivalences $\{\langle \mathtt{X} \rangle | \mathtt{X} \in \mathfrak{U}_{\mathtt{A}} \cup \{\mathtt{A}\}\}$ forme une partition $2^{\mathtt{A}}$. On a donc :

$$|2^{\mathtt{A}}| = \sum_{\mathtt{X} \in \mathfrak{U}_{\mathtt{A}}} |\langle \mathtt{X} \rangle| + |\langle \mathtt{A} \rangle|$$

Ce qui donne :

$$|\langle \mathtt{A} \rangle| = |2^{\mathtt{A}}| - \sum_{\mathtt{X} \in \mathfrak{U}_{\mathtt{A}}} |\langle \mathtt{X} \rangle|$$

En divisant par $|2^{\mathtt{A}}|$ on obtient :

$$\begin{aligned} \frac{|\langle \mathtt{A} \rangle|}{|2^{\mathtt{A}}|} &= 1 - \sum_{\mathtt{X} \in \mathfrak{U}_{\mathtt{A}}} \frac{|\langle \mathtt{X} \rangle|}{|2^{\mathtt{A}}|} \\ \gamma(\mathtt{C}) &= 1 - \sum_{\mathtt{X} \in \mathfrak{U}_{\mathtt{A}}} \gamma(\mathtt{X}, \mathtt{X}') 2^{|\mathtt{X}| - |\mathtt{A}|} \end{aligned}$$

□

Cette propriété a deux conséquences. Tout d'abord, la stabilité d'un concept dépend de celle de ses sous-concepts, et, plus ces derniers sont stables et lui ressemblent du point de vue de leur extension, moins il est stable. La deuxième conséquence est d'ordre pratique. Le calcul de la stabilité est un problème #P-complet [Kuznetsov, 2007]. Cette classe rassemble les problèmes d'énumération de solutions à des problèmes de classe NP. Néanmoins, une fois le treillis des concepts connu, il est alors possible de calculer plus rapidement la stabilité d'un concept à l'aide d'un algorithme de parcours ascendant du treillis.

3.3 Analyse qualitative de la stabilité

3.3.1 Stabilité et cohésion

Pour Kuznetsov *et al.* [Kuznetsov et al., 2007], l'idée sous-jacente d'un concept stable est que son intension ne dépend pas trop d'objets particuliers de son extension. La stabilité est de ce fait bien adaptée à l'analyse des communautés sociales par des treillis de concepts. Les communautés sociales sont des groupes d'individus caractérisés par des intérêts ou objectifs communs. Elles peuvent être explorées et visualisées par AFC (cf section 5.2). Par exemple, un groupe de chercheurs travaillant sur des domaines de recherche communs peut être identifié comme une communauté sociale bien définie. On peut modéliser ce problème par un contexte formel dans lequel les chercheurs sont des objets et les problèmes de recherche qu'ils étudient des attributs. Les concepts obtenus dans le treillis correspondant définissent ainsi des communautés de chercheurs caractérisés par des objectifs partagés. Toujours selon Kuznetsov [Kuznetsov et al., 2007], une communauté doit posséder une cohésion interne suffisante : le concept du treillis représentant une communauté à forte cohésion interne reste un concept même si quelques membres quittent la communauté. Un tel concept est par conséquent résistant au bruit et ne se décompose pas en sous-concepts si on supprime quelques objets de son extension. Les concepts les plus stables permettent donc d'identifier des groupes d'objets intéressants constitutifs d'une classe à forte cohésion interne.

3.3.2 Les concepts stables sont intéressants

L'AFC a été utilisée avec succès pour l'identification de communautés dans des réseaux sociaux[Kuznetsov et al., 2007]. Néanmoins, l'AFC est confrontée à une croissance potentiellement exponentielle de la taille des treillis, fonction du nombre d'objets et d'attributs des contextes formels. Il s'agit d'une conséquence problématique pour l'étude de réseaux sociaux de grande taille.

Pour réduire la complexité des treillis de grande taille, on fait appel à des mesures d'intérêt des concepts formels, comme par exemple le support. Pour autant, filtrer les concepts sur leur support repose sur l'assomption que les connaissances potentiellement intéressantes sont représentées par des phénomènes fréquents. Cependant, dans de nombreux domaines, les chercheurs se focalisent sur l'étude de phénomènes rares. On peut citer par exemple la recherche de mutations rares en génétique ou la survenue d'effets indésirables médicamenteux rares, mais sévères, dans le domaine de la pharmacovigilance. Cette question a par ailleurs déjà été soulevée par les travaux de Szathmary [Szathmary et al., 2007] sur l'extraction de règles d'association de faible support mais de confiance élevée.

Une approche fondée sur la stabilité porte un regard différent sur l'intérêt des concepts formels. Dans leurs travaux sur les communautés épistémologiques, Roth *et al.* [Roth et al., 2006, Roth, 2006] montrent, grâce à l'AFC, comment s'organisent des communautés de chercheurs autour de domaines communs. Pour ce faire, ils construisent des contextes dont les objets sont des auteurs et les attributs des mots clefs de leurs publications. Si les icebergs parviennent à identifier des communautés importantes rassemblées autour de problèmes fréquemment étudiés, ils échouent à isoler de petites communautés « émergentes », soudées autour de problèmes nouveaux. L'utilisation de la stabilité leur a permis de repérer ces dernières. En éliminant le bruit introduit par des concepts issus d'une co-occurence de mots-clés probablement fortuite, Roth *et al.* détectent de petites communautés à forte cohésion représentées par des concepts stables de faible support. En cela, la stabilité donne un pouvoir de discrimination supplémentaire pour différencier des concepts dans des intervalles de support étroits.

De plus, la stabilité permet de repérer des concepts fréquents fortement liés à la présence d'objets singuliers dans leur extension. Cette propriété est illustrée par la figure 3.2. Ce treillis est composé de deux concepts très semblables :

- $\mathtt{C_1} = (\{\mathtt{g_1}, ..., \mathtt{g_{n-1}}, \mathtt{g_n}\}, \{\mathtt{m_1}\})$
- $\mathtt{C_2} = (\{\mathtt{g_1}, ..., \mathtt{g_{n-1}}\}, \{\mathtt{m_1}, \mathtt{m_2}\})$.

Ils ne diffèrent que par l'objet $\mathtt{g_n}$ qui ne possède pas l'attribut m_2. En considérant n suffisamment élevé, tous deux ont un support proche de $\mathtt{1}$. Mais quelle que soit la valeur de $\mathtt{n}$ la stabilité de $\mathtt{C_1}$ vaut $\frac{1}{2}$ alors que celle de $\mathtt{C_2}$ vaut $\mathtt{1}$. En termes de communauté, le groupe des individus $\{\mathtt{g_1}, ..., \mathtt{g_n}\}$ n'a pas de cohésion interne suffisante et n'aurait probablement pas

d'existence réelle. Dans ce cas précis, la stabilité s'avère insensible à la fréquence du phénomène étudié, ce qui évidemment n'est pas le cas du support. Cette propriété a des conséquences très intéressantes lorsque l'observation d'un phénomène ne dépend pas du volume des données collectées pour l'étudier. Ce cas de figure peut se retrouver fréquemment en présence d'un système d'information exhaustif.

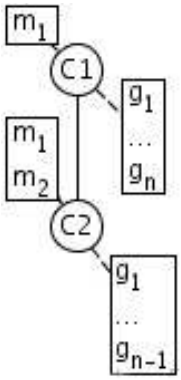

FIG. 3.2 – C1 : un concept fréquent instable

Au final, la stabilité, associée au support, permet de distinguer deux types de concepts potentiellement intéressants :

- les *concepts stables et rares* qui ont une stabilité élevée et un support faible.
- les *concepts instables et fréquents* qui ont une stabilité faible et un support élevé.

3.3.3 Les concepts stables sont monothétiques

Comme cela est décrit plus haut, la stabilité d'un concept est liée à la présence d'exceptions dans les objets qui le caractérisent. En classification, la présence d'exceptions dans les instances d'une classe peut être discutée du point de vue de son caractère *monothétique* ou *polythéthique* [Sokal, 1968, bailey, 1973, Lerman, 1970].

Définition 40 (Monothétisme)**.** *Une classe d'individus* `C` *est dite* monothétique *si et seulement si il existe un ensemble d'attributs* `Att` *qui déterminent l'appartenance d'un individu à cette classe. (*`Att` *est un ensemble de conditions nécessaires et suffisantes)*

Définition 41 (Polythétisme)**.** *Étant donné un ensemble d'attributs* `Att`*, une classe* `C` *est dite* polythétique *si :*

- *chaque instance de* `C` *possède un nombre important, pas nécessairement fixe, d'attributs de* `Att`*.*
- *Chaque attribut de* `Att` *est possédé par un nombre important, pas nécessairement fixe, d'instances de* `C`*.*

– *Il n'y a pas nécessairement d'attribut de* `Att` *qui soit partagé par toutes les instances de* `C`

En s'appuyant sur le fait qu'un concept stable est résistant au bruit, et qu'il ne se décompose pas si on retire certains objets de son extension, plus un concept est stable, moins il représente d'individus exceptionnels, et plus il a de légitimité à représenter des groupes à forte cohésion interne comme les communautés sociales. Intuitivement, ceci laisse supposer que les objets des concepts stables appartiennent à des classes plutôt monothétiques. Les concepts instables, quant à eux, présentent un caractère plus polythétique : s'ils sont assez similaires, leurs objets n'en présentent pas moins des différences qui permettent de décomposer un concept instable en de nombreux sous-concepts. Ainsi, un concept instable peut être vu comme la réunion d'un grand nombre de sous-classes, alors qu'un concept stable est plus difficilement décomposable.

La figure 3.3 illustre ce point de vue. A gauche, le contexte formel fait clairement apparaître deux classes d'objets : $\mathtt{c_1} = \{1, 2, 3, 4\}$ et $\mathtt{c_2} = \{5, 6, 7, 8\}$. Pris deux à deux, les objets de la classe $\mathtt{c_1}$ ont systématiquement trois attributs communs parmi $\{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}\}$, mais aucun de ces attributs n'est partagé par tous. Ceci confère un caractère polythétique à $\mathtt{c_1}$. Cette classe donne naissance à 14 concepts dans le treillis de droite. Tous ces concepts ont une stabilité très basse, en particulier les plus grands d'entre eux (étiquetés par $\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}$). Aucun d'entre eux n'explique à lui seul la similarité qui caractérise les objets de la classe $\mathtt{c_1}$. En revanche la classe $\mathtt{c_2}$ présente un caractère plus monothétique. Elle est représentée dans le treillis par seulement trois concepts (étiquetés $\{\mathtt{e}, \mathtt{f}\}$, $\{\mathtt{h}\}$ et $\{\mathtt{g}\}$) à la stabilité plus élevée. La classe $\mathtt{c_2}$ peut être décomposée en deux sous-classes, d'objets identiques au regard de leurs attributs, $\{5, 6\}$ et $\{7, 8\}$. Ces sous-classes correspondent aux concepts, étiquetés par $\mathtt{g}$ et $\mathtt{h}$, qui ont la stabilité la plus élevée dans le treillis. La partie gauche du treillis est sensible au bruit : si l'on supprime l'un des objets $\{1, 2, 3, 4\}$, la structure sera substantiellement modifiée. A l'opposé, la parti droite du treillis est plus robuste. La suppression d'un objet parmi $\{5, 6, 7, 8\}$ ne va pas modifier sa structure. Au final ceci confirme que les concepts les plus stables représentent des familles monothétiques d'objets, alors que les moins stables représentent des familles polythétiques.

3.4 Conclusion

L'utilisation de l'AFC dans la phase d'interprétation du processus d'ECD a longtemps souffert des limites de manipulation des treillis de concepts volumineux. L'introduction des icebergs de concepts a permis de franchir en partie cette limite. Toutefois, l'approche voulant que des connaissances intéressantes correspondent à des phénomènes fréquents dans les données n'est pas toujours adaptée. Nous pensons que la stabilité présente une alternative séduisante pour la mesure de l'intérêt des concepts. Cette conviction repose sur l'utilité potentielle de cette mesure en présence de données bruitées. Par ailleurs, elle permet d'envisager la découverte de

	a	b	c	d	e	f	g	h
1	×	×	×					
2	×	×		×				
3	×		×	×				
4		×	×	×				
5					×	×	×	
6					×	×	×	
7					×	×		×
8					×	×		×

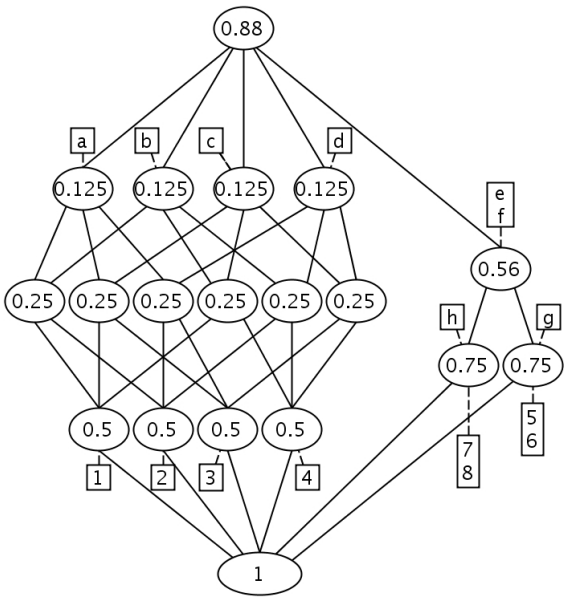

FIG. 3.3 – Concepts monothétiques et polythétiques

connaissances portant sur des phénomènes rares.

Chapitre 4

Le contexte médical

Sommaire

4.1 Introduction

Dans la plupart des pays occidentaux, les dépenses de santé ne cessent d'augmenter. Les facteurs de cette augmentation sont le vieillissement de la population, l'amélioration technologique des soins, l'information du public et l'accès facilité au système de santé, l'élargissement des actions de santé... Parallèlement, l'utilisation de thérapeutiques inappropriées ou de techniques d'investigation inutiles ou dangereuses est difficile à contrôler et source de disparités importantes dans les pratiques médicales. Une pression de plus en plus importante pèse donc sur les systèmes de soins pour concilier qualité des soins et maîtrise des dépenses. Cette pression doit les amener à une utilisation plus efficace des connaissances médicales.

Apparue dans les années 1990, la médecine factuelle, renforcée par les guides de bonnes pratiques cliniques, a modifié peu à peu l'exercice en fondant le raisonnement médical sur des connaissances actualisées issues de la recherche clinique. Mais le chemin peut être long de la publication scientifique d'une découverte à son application sur le terrain. De plus, cette mise en pratique peut se heurter à des problèmes organisationnels, surtout lorsque la prise en charge des patients nécessite une approche pluridisciplinaire, impliquant l'intervention de plusieurs acteurs. Ainsi, une connaissance approfondie du fonctionnement du système et des processus

de soins est indispensable à l'amélioration globale de la qualité et de l'efficience des soins. Cette connaissance peut être acquise par une observation des processus de soins et transformée en un savoir organisationnel sur lequel pourront s'appuyer les stratégies de planification et d'organisation des soins. L'observation du fonctionnement réel du système de soins nécessite de disposer de systèmes d'information adaptés au suivi des *trajectoires de soins* de patients dans un environnement donné. Si de nombreux systèmes d'information sanitaires ont été mis en place dans les pays développés, et particulièrement en France, ils n'ont pas été conçus spécifiquement dans ce but. Parmi eux, les Systèmes de Classification de Malades (SCM) produisent malgré tout d'énormes volumes de données, souvent sous-exploitées, dont les méthodes d'ECD pourraient avantageusement tirer parti.

Après une présentation de la médecine factuelle, ce chapitre décrit les problèmes d'apprentissage et de savoir organisationnel en médecine. La deuxième section définit la notion de trajectoire de soins et présente les SCM. Enfin, la dernière section s'attarde plus spécifiquement sur le PMSI, un SCM français, qui sera à la source de l'analyse des trajectoires de soins par treillis de concepts, objet du chapitre suivant.

4.2 Du savoir individuel au savoir organisationnel

4.2.1 Médecine factuelle et guides de bonnes pratiques cliniques

La médecine factuelle (Evidence-based medecine ou EBM), née au Canada au début des années 1980, est un processus de recherche, évaluation et utilisation systématique de données de recherches communautaires et cliniques contemporaines comme fondement des décisions cliniques. L'objectif de cette démarche est d'assurer la meilleure gestion possible de la santé et de la maladie ainsi que de leurs déterminants au niveau communautaire [Jenicek and Stachenko, 2003]. Si l'EBM a modifié la pratique de la médecine, ce processus est loin d'être achevé et les découvertes de la recherche mettent parfois longtemps avant d'être appliquées sur le terrain. En particulier, la prolifération de la littérature scientifique pose ses propres problèmes de diffusion des connaissances. Le rôle des guides de pratique clinique est précisément d'assister les médecins et leurs patients dans la prise de décision la plus adaptée aux circonstances cliniques particulières. Énoncés la plupart du temps par des sociétés savantes ou l'autorité sanitaire, les guides de bonne pratique sont en cela une traduction du savoir, issu de la recherche, en action.

Les cliniciens, les décideurs et les assureurs voient dans les guides de bonne pratique un outil pour à la fois diminuer la variabilité des pratiques, contrôler les coûts, et améliorer la qualité des soins. Toutefois, le développement d'une bonne recommandation n'assure pas forcément son utilisation en pratique [Feder et al., 1999], même si l'utilisation des technologies de l'information peut accélérer sa diffusion [Fox et al., 1998, D'Aquin et al., 2004, Dufour et al., 2006].

En se heurtant à la réalité opérationnelle des systèmes de soins, les recommandations peinent à s'imposer dans le cadre d'une prise en charge multidisciplinaire, impliquant plusieurs acteurs voire plusieurs institutions. Pour Stefanelli [Stefanelli, 2001], si l'amélioration de la qualité des soins et la diffusion des bonnes pratiques nécessitent un apprentissage individuel de la part des soignants, elles passent également par un apprentissage aux niveau des organismes de soins eux-mêmes.

4.2.2 Apprentissage et savoir organisationnel en médecine

Un hôpital n'est pas seulement une collection d'individus : il se compose de multiples interactions entre plusieurs équipes, chacune dotée de compétences, de connaissances et de techniques spécialisées. L'exécution des tâches de soins requiert de la part des ces communautés une intégration de leurs spécificités pour créer une nouvelle perspective partagée, c'est-à-dire du savoir au niveau organisationnel. C'est ce savoir organisationnel qui pourrait permettre, par exemple, d'améliorer la prise en charge au long cours de patients atteints de maladie chronique en redéfinissant ou en adaptant les processus de soins. Argyris et Schön [Argyris and Schön, 2002] modélisent l'apprentissage du savoir organisationnel comme un processus avec une double boucle de contrôle. La figure 4.1 en montre une adaptation au contexte médical proposée par Stefanelli [Stefanelli, 2001]. La première boucle (simple) implique que la gestion des processus de soins est décidée en fonction de la comparaison entre les résultats attendus (objectifs) et obtenus (issues). Ils ne peuvent être mesurés qu'en définissant des indicateurs (valeurs) d'efficacité des soins et d'efficience d'utilisation des ressources. Ces indicateurs peuvent prendre la forme d'indices d'efficacité clinique, de sécurité des soins ou de qualité de vie des patients. Quand les résultats attendus et observés convergent, seule la boucle simple d'apprentissage est mise en jeu. Lorsqu'ils divergent, le processus de réorganisation (boucle double) s'enclenche jusqu'à ce qu'une stratégie améliorant la performance globale soit identifiée. Le système réactive alors la boucle simple.

La capacité d'apprentissage des organisations dépend de leur faculté à cerner leur contexte d'apprentissage ou, en d'autres termes, à identifier quand et comment elles doivent apprendre, quand et pourquoi elles n'apprennent pas, et quels sont leurs mécanismes d'adaptation. C'est ce que l'on peut qualifier de méta-apprentissage. Pour cela, elles doivent mettre en place un processus de gestion des connaissances capable d'entretenir leur cycle d'apprentissage. Boisot [Boisot, 1995] modélise le méta-apprentissage comme un cycle de quatre opérations, représenté sur la figure 4.2, dans lequel le savoir est qualifié par deux propriétés : son degré de *représentation* et son degré d'*utilisation*.

Le degré de représentation distingue le savoir tacite du savoir explicite. Le degré d'utilisation distingue le savoir latent du savoir manifeste. L'*externalisation* transforme un savoir tacite

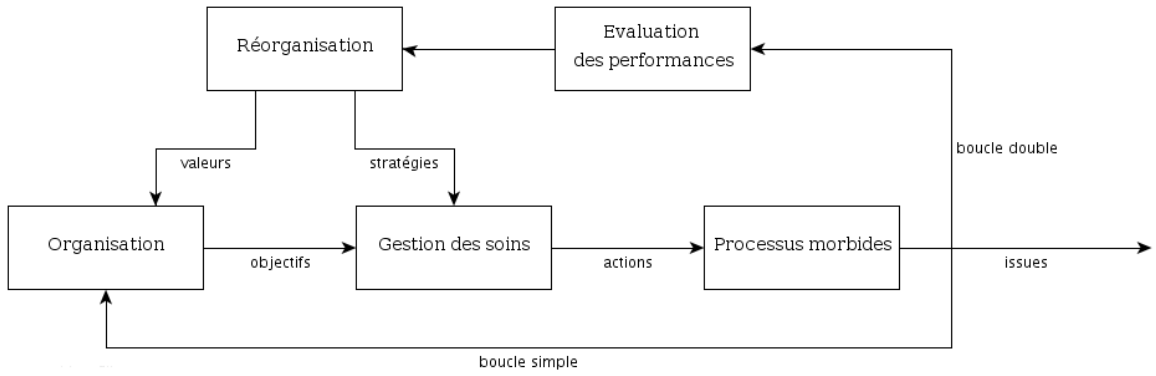

FIG. 4.1 – Processus d'apprentissage organisationnel d'une structure de soins (adapté de [Stefanelli, 2001]).

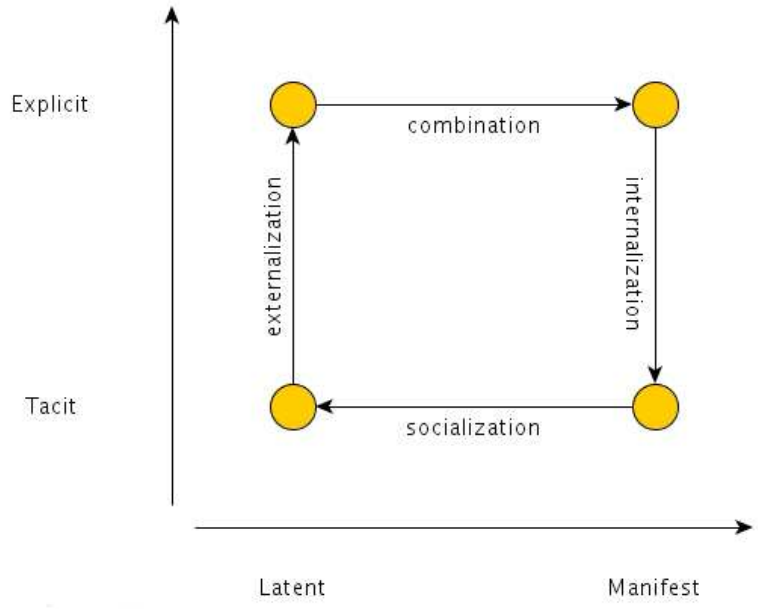

FIG. 4.2 – Le cycle de méta-apprentissage organisationnel par Boisot.

en un savoir explicite à travers le développement de modèles, protocoles et autres recommandations. La *combinaison* reconfigure et recombine les connaissances explicites pour établir de nouvelles connaissances. L'*internalisation* est le processus de répétition d'une tâche par des individus en appliquant des connaissances explicites pour en faire des connaissances tacites. Enfin, la *socialisation* est le transfert du savoir tacite des praticiens expérimentés vers les novices par l'observation et l'imitation. On peut resituer les guides de pratique clinique dans le cycle d'apprentissage : ils découlent du processus de combinaison et initient le processus d'internalisation. Parfois, un évènement que Stefanelli [Stefanelli, 2001] compare à une mutation se produit : l'application idiosyncrasique d'une connaissance entraîne des conséquences inattendues, potentiellement utiles et non partagées par la communauté. Les méthodes d'extraction des connaissances à partir de données peuvent supporter ces découvertes. De plus, les outils de représentation des connaissances offrent la possibilité de les rendre explicites et contribuer ainsi au processus d'externalisation.

Pour Stefanelli, l'apprentissage des organismes de soins passe par l'exploitation des Technologies de l'Information et de la Communication (TIC) pour une gestion des connaissances, de l'expertise et des compétences de trois différents domaines : médical, organisationnel et technique. Par connaissance médicale, on entend toute représentation d'un ensemble d'activités résultant de l'interaction entre patients et soignants. Les professionnels travaillent au sein de structures dont dépendent leurs moyens financiers et matériels. Leurs comportements sont dictés par des règlements, des lois, mais aussi des règles et pratiques non écrites. La compréhension sociologiques de ces mécanismes complexes est indispensable à l'apprentissage organisationnel. Les TIC sont à la fois un support et un vecteur de connaissance à travers lesquels les personnes doivent pouvoir mieux communiquer et mieux collaborer. Néanmoins, en raison de besoins et de contextes différents, plusieurs points de vue peuvent coexister et représenter une même réalité. Il faut s'employer à construire des outils qui réduisent les divergences conceptuelles et terminologiques pour établir une base de communication entre individus et d'inter-opérabilité entre systèmes. Sur ce point, les ontologies peuvent constituer un cadre de travail unifié.

4.2.3 Programmes de soins

Les programmes de soins figurent en bonne place parmi les outils susceptibles d'accélérer le cycle d'apprentissage des organisations en médecine. Le concept de programme de soins est né aux États-Unis en 1987 [DeZell et al., 1987]. Il est actuellement largement répandu, en particulier au Royaume Uni et aux États-Unis. Il regroupe cependant des réalités différentes comme en témoignent ses multiples synonymes en anglais : clinical pathway, critical pathway, care pathway, integrated care pathway ou care map. Une étude menée par De Luc *et al.* [de Luc et al., 2001] retrouve même 17 termes recouvrant ce concept parmi lesquels : clinical progression, clinical

outcomes, care protocol, care profile, collaborative care plan, etc. Dans le Medical Subject Headings[1] (MeSH), le concept de critical pathway a été introduit en 1996 et défini comme : « Schedules of medical and nursing procedures, including diagnostic tests, medications, and consultations designed to effect an efficient, coordinated program of treatment », c'est-à-dire un programme de procédures médicales et infirmières, incluant des tests diagnostiques, des médicaments, et des consultations, conçu pour délivrer un traitement efficace et coordonné.

A partir de 283 articles de la littérature médicale, [Bleser et al., 2006] proposent de redéfinir le concept de programme de soins. Les auteurs identifient les termes les plus fréquents et les classent en 16 catégories, reproduites dans le tableau 4.1.

Catégorie	n[a]
Groupe homogène de patients	45
Équipe multidisciplinaire	36
Échelle de temps	30
Inventaire des actions	29
Gestion des soins centrée sur le patient	17
Efficience des soins	15
Standardisation des soins	12
Séquence dans la programme de soins	9
Analyse de variabilité	9
Médecine factuelle	8
Amélioration de la qualité des soins	6
Élément du dossier patient	6
Guides de bonne pratique	4
Éducation du patient et des soignants	4
Moyen de communication	4
Recueil de données	3

[a]occurrences des termes se rattachant à chaque catégorie

TAB. 4.1 – Caractéristiques des programmes de soins.

La première constatation est que les programmes de soins s'appliquent à des patients présentant un certain nombre de similarités, et en premier lieu une pathologie commune. La deuxième souligne le caractère multidisciplinaire des programmes de soins qui s'inscrit dans l'espace. Les contacts d'un patient avec le système de soins sont souvent multiples, surtout s'il est question de pathologie chronique, et font intervenir des acteurs différents : médecins généralistes, spécialistes, hôpitaux, infirmières, kinésithérapeutes... D'autre part le programme de soins s'inscrit dans une échelle de temps. C'est un aspect particulièrement difficile à modéliser : les limites temporelles des processus pathologiques sont parfois floues. S'intéresse-t-on à une maladie qui

[1]http ://www.nlm.nih.gov/mesh/MBrowser.html

peut durer des années ou seulement à un épisode particulier correspondant à une phase d'accutisation ? Quand débute une maladie : au moment du diagnostic, ou des premiers symptômes ? Définir le cadre temporel du programme de soins doit permettre de faire l'inventaire des actions entreprises dans la prise en charge du patient, qui se déroule d'une manière séquentielle. L'ordre des interventions revêt une importance particulière, comme celui des états de santé successifs du patient. Enfin, pour la plupart des professionnels, un programme optimal minimise les coûts tout en maximisant la qualité des soins. Pour cela, il s'appuie sur des guides de bonne pratique et peut constituer ou puiser des éléments dans le dossier médical, ce qui implique un recueil de données et en fait un instrument de communication.

Au final, Bleser *et al.* [Bleser et al., 2006] décrivent le programme de soins comme une méthode de gestion des soins centrée sur le patient, s'appliquant un des groupes bien définis de patients sur une durée bien délimitée. Elle explicite les objectifs et les éléments clefs de la prise en charge sur la base de l'EBM, en facilitant la communication, la coordination des rôles et le séquencement des interventions d'équipes multidisciplinaires. Son but ultime est l'amélioration de la qualité des soins, la réduction des risques, une satisfaction accrue des patients et l'utilisation efficiente des ressources. Pour cela, elle met en place des instruments de mesure qui permettent de documenter et d'évaluer la variabilité des pratiques.

Bien que théoriquement fondés sur des guides de bonne pratique, les programmes de soins en sont une adaptation à des spécificités locales. Que ce soit dans leur conception ou pour leur évaluation, ils reposent sur une connaissance approfondie du tissu sanitaire local et du parcours des patients dans le système de soins. En particulier, les systèmes d'information médicaux devraient être capables d'établir et d'analyser des profils de ce que l'on appelle les *trajectoires de soins*.

4.3 Trajectoires de soins et systèmes de classification de malades

Á l'instar du concept de programme de soins, celui de trajectoire de soins recouvre de multiples réalités. Pour Riou et Jarno [Riou and Jarno, 2000], il importe d'abord de différencier les notions de « filière de soins » et de « trajectoire de soins », deux termes utilisés pour parler des déplacement d'un patient dans le système de soins. Derrière le celui de filière se dessine l'idée d'une prédétermination par les professionnels, alors que le terme de trajectoire apparaît plus neutre, laissant une marge de liberté au patient lui-même. D'autre part, le langage officiel a restreint l'utilisation du terme de filière à un parcours de soins passant obligatoirement par le médecin généraliste.

Bien que la représentation des trajectoires de soins ne soit pas normée, Riou et Jarno lui distinguent trois éléments indispensables :

- la population concernée ; il peut s'agir du tout venant, ou bien de patients sélectionnés par exemple sur leur âge ou selon une pathologie précise ;
- les bornes ; elles peuvent être physiques (un espace géographique, un ensemble de soignants), ou bien temporelles (une année calendaire, la durée d'une pathologie)
- le contenu ; il s'agit de l'ensemble des éléments de parcours observés entre les bornes de la trajectoire.

Il est difficile de définir plus précisemment la notion de trajectoire de soins sans perdre en généralité. De fait, la caractérisation de la population, des bornes et du contenu des trajectoires de soins dépendent de l'objectif poursuivi par leur modélisation [Riou and Jarno, 2000, Naiditch, 1993] : ajustement du risque pour le calcul d'une prime d'assurance, outil de paiement des professionnels, gestion des ressources, analyse de la qualité des pratiques professionnelles.

Selon Riou et Jarno, cette modélisation s'appuie sur la construction de typologies à partir du regroupement de trajectoires observées, ce qui exclue une modélisation à priori à dires d'experts. L'importance de cette approche est également soulevée par Grémy [Grémy, 1997] pour qui, chaque patient étant singulier, la trajectoire qu'il suit est unique. Grémy voit ainsi les filières de soins comme « des regroupements de trajectoires fréquentes et/ou typiques et de fait susceptibles d'une certaine forme de modélisation ».

Les modalités de représentation de trajectoires peuvent être source de biais d'interprétation qui dépendent en grande partie de la définition des bornes d'observation du parcours de soins. Pour Riou et Jarno [Riou and Jarno, 2000], comme pour Naiditch [Naiditch, 2000], la définition de telles bornes passe par le découpage de la trajectoire en « épisodes » de soins. Naiditch propose une définition synthétique de l'épisode de soins à partir des travaux de Hornbrook [Hornbrook and A., 1985, Hornbrook, 1993] : « Période (avec un début et une fin théoriquement identifiables) durant laquelle un problème de santé, une pathologie spécifique, un symptôme perçu par un patient, ou un traitement, est présent et donne lieu à un ensemble de prestations délivrées par un ensemble de professionnels ou d'institutions ». Il s'agit là encore d'une définition très large qui a l'avantage d'englober différents contenus (des diagnostics, des symptômes, des traitements) et différents points de vue (celui des patients, des professionnels, des institutions ou encore des régulateurs).

L'analyse des trajectoires de soins à l'échelle d'une population et leur regroupement en classes homogènes présentent des liens étroits avec le concept de système de classification de malades. (SCM). Hornbrook [Hornbrook, 1982] en donne la définition suivante : ce sont des systèmes qui réduisent l'infinie variété des malades en groupes de malades semblables entre eux au regard de plusieurs caractéristiques ou variables explicatives. Ces regroupements sont spécifiques de l'objectif précis que l'on vise :

- description clinique (groupes iso-symptômes, iso-diagnostics, iso-syndrômes),

- résultats de soins (groupes iso-résultats),
- consommation de ressource (groupes iso-ressources)

L'un des tous premiers SCM fut développé à partir des années 1960 par Fetter [Fetter et al., 1980] : les Diagnosis Related Groups (DRG). Son objectif était de mettre en place une méthode efficace de contrôle de la qualité des soins et d'utilisation des ressources hospitalières en définissant des groupes de prise en charge comparable. Les DRG sont un algorithme de classement des séjours hospitaliers à partir d'un ensemble minimal d'informations constitué :

- de diagnostics,
- de procédures médicales et chirurgicales,
- de données administratives : services médicaux fréquentés, durées de séjour,
- de données démographiques : âge du patient, sexe.

Les DRG ont donné naissance à de nombreuses adaptations dans le monde. En France, il s'agit du Programme de Médicalisation des Systèmes d'Information (PMSI), mis en place au milieu des années 1980, et qui joue aujourd'hui un rôle majeur dans l'analyse de l'activité et le financement des hôpitaux. De nombreux SCM sont construits sur le même principe : le regroupement des séjours résulte d'une analyse de la variance du coût par segmentation. En l'occurrence, la méthode utilisée par les concepteurs des DRG est l'Automatic Interaction Detection (AID) [Green, 1978]. Pour assurer un plus grande cohérence clinique, la classification obtenue a été modifiée à dire d'expert pour rapprocher les séjours causés par des problèmes médicaux semblables. Autre point commun, ces SCM se limitent à l'analyse d'un séjour hospitalier unique.

D'autres SCM ont été élaborés pour répondre à des objectifs de classification différents. Leur critère de classification est le degré de sévérité de la maladie, comme le Disease Staging [Gonnella et al., 1984], ou encore un indicateur pronostique (taux de mortalité, taux de morbidité) comme les Medigroups [Brewster et al., 1985]. Là encore, ces systèmes se limitent à l'analyse d'un unique contact du patient avec le système de soins.

Pour rendre compte de la globalité du patient et intégrer une plus grande partie de sa trajectoire de soins, des équipes se sont penchées sur des SCM orientés patients et épisodes de soins. Le plus ancien est le système ACG (Ambulatory Care Groups) [Weiner et al., 1991]. Ils suivent une approche longitudinale et plus holistique en tenant compte de plusieurs contacts patients/soignant constituant un épisode de soins. La problématique des systèmes de classification d'épisodes de soins présentent au moins deux spécificités :

- comment tenir compte des polypathologies ?
- comment définir les bornes temporelles d'un épisode, en particulier dans le cas des maladies chroniques ?

En ce qui concerne le problème des polypathologies, alors que dans les SCM de type DRG, la règle qui prévaut est « un séjour = un groupe », les systèmes de classification d'épisodes peuvent classer un même patient dans plusieurs groupes. Pour répondre à la question des bornes de l'épisode plusieurs solutions ont été envisagées. Tout d'abord, la définition du début correspond généralement à un ou plusieurs diagnostics marqueurs. Ensuite, soit les épisodes ont chacun une longueur fixe définie à priori suivant le problème médical et le problème de la fin ne se pose pas, soit on admet que l'épisode à une longueur variable (finie). Dans ce cas la notion de fenêtre vide est utilisée : tout épisode ne comportant aucun nouvel évènement après le dernier contact enregistré est considéré clos.

Les SCM représentent une approche particulièrement intéressante pour analyser les trajectoires de soins à un niveau populationnel. Bien qu'ils constituent déjà un ensemble de systèmes de classification des trajectoires, ils présentent un certains nombre de limites. Tout d'abord, les systèmes de type DRG sont restreints à l'étude d'un seul évènement de santé, ne reflétant qu'une infime partie de la trajectoire de soin, ce qui est problématique dans le suivi des pathologies chroniques. Mais surtout, dans une grande majorité, leur construction à été guidée par un objectif précis, souvent médico-économique : mesure de l'utilisation des ressources, calcul de risque pour l'ajustement de primes d'assurance, outil de paiement. Ainsi, leur principe de construction repose souvent sur l'analyse de variance d'une variable à expliquer : par exemple le coût ou un score d'état de santé. Néanmoins, lorsque la nécessité se fait sentir de disposer d'une information « médicalisée »,cette analyse est par ailleurs raffinée par des experts pour constituer des groupes cliniquement cohérents. Enfin, ils sont indissociables de la mise en place d'un système d'information dédié qui les alimente. Le déploiement de tels systèmes est particulièrement coûteux, non seulement en termes financiers, mais aussi dans les changements qu'il entraîne dans les habitudes de travail. Ces coûts ne doivent pas être négligés, d'autant plus que la maîtrise des dépenses fait souvent partie des finalités d'un SCM. L'ensemble de ces limites confère une certaine rigidité aux SCM. Néanmoins, ils produisent chaque année d'énormes quantités d'informations susceptibles d'être réutilisées pour l'analyse des trajectoires de soins.

4.4 ECD, PMSI et trajectoires de soins

Les limites posées par les SCM, la grande quantité d'information qu'ils recèlent, et la nécessité de disposer d'outils d'apprentissage organisationnel qui permettent d'analyser les trajectoires de soins définissent un champ idéal d'application de l'ECD. Il y a là la possibilité de bâtir des systèmes de classification et de représentation des connaissances ad hoc, souples et à moindre coût à partir de données existantes. Notre objectif est donc d'évaluer le potentiel de l'AFC dans ce contexte. Pour cela, nous nous appuyons sur les données issues du PMSI.

En France, plusieurs auteurs se sont penchés sur l'analyse de trajectoires de soins à partir de données issues des systèmes d'information en santé. Dans une approche supervisée, Le Duff *et al.* [Duff et al., 2001] ont utilisé des arbres de décision pour créer des profils de consommation et entraîné un réseau de neurones pour détecter des patients souffrant de lombalgie. Les données utilisées provenaient d'une base alimentée par les systèmes d'information de l'assurance maladie, du PMSI et de différents professionnels de santé. Dans [Dart et al., 2003], Dart *et al.* ont de leur coté mis au point un outil d'analyse des trajectoires intra-hospitalières (inter-services) à partir de la recherche de motifs séquentiels fréquents et de règles d'associations. Là encore, les auteurs ont recours aux données du PMSI. Elghazel [Elghazel, 2007] a quant à lui élaboré une méthode classification de séquences temporelles par modèles de mélange markoviens, également mise en œuvre sur des données du PMSI.

Le PMSI est un SCM largement utilisé en France. Il couvre l'ensemble des hospitalisations en soins aigus sur le territoire national. Le PMSI est un SCM iso-ressources qui classe les hospitalisations dans des groupes présentant une double cohérence sur le plan clinique et celui de la mobilisation des ressources hospitalières : les groupes homogènes de malades (GHM). Le terme de *malade* est ici plutôt inapproprié puisque le PMSI ne traite à chaque fois qu'un seul séjour hospitalier.

Dans ce système, chaque séjour donne lieu à la réalisation d'un Résumé Standardisé de Sortie (RSS), constitué d'un petit nombre d'informations administratives (âge, durée de séjour...) et médicales (diagnostic principal, comorbidités, actes médicaux et chirurgicaux) relatives au patient pendant ce séjour. L'obtention de ces informations conditionne le fonctionnement du dispositif PMSI dans toutes ses implications et utilisations, qu'elles soient d'ordre budgétaire ou autre.

Le RSS est divisé en Résumés d'Unité Médicale (RUM). On désigne par unité médicale un ensemble individualisé de moyens assurant des soins à des malades hospitalisés, repéré par un code spécifique dans une nomenclature déterminée par l'établissement. L'information médicale contenue dans le RUM obéit à des règles de hiérarchisation bien définies. Ainsi, après le passage d'un patient dans une unité médicale, le médecin responsable détermine quel motif de soins a mobilisé l'essentiel de l'effort médical et soignant. Celui-ci devient le diagnostic principal (DP). Il doit être obligatoirement renseigné à chaque séjour en unité médicale. Il existe d'autres types de diagnostics, dont le renseignement, facultatif, permet de mieux caractériser le séjour d'un patient :

- le Diagnostic Relié (DR) a pour rôle, en association avec le DP et lorsque celui-ci n'y suffit pas, de rendre compte de la pathologie du patient.
- les Diagnostics Associés Significatifs (DAS) sont d'autres morbidités ayant donné lieu à une prise en charge diagnostique ou thérapeutique supplémentaire au cours du séjour. Ils

peuvent influer sur la classification du séjour en GHM comme un marqueur de sévérité. Les informations médicales sont des informations codées :

- au moyen de la Classification statistique Internationale des Maladies et problèmes de santé connexes (CIM10) pour les diagnostics,
- au moyen de la Classification Commune des Actes Médicaux pour les actes médicaux et chirurgicaux.

A l'issue du séjour hospitalier, le RSS d'un patient peut être classé dans un GHM grâce à un algorithme de groupage qui tient compte des durées de séjours, de l'âge, du sexe, des diagnostics et actes médicaux collectés durant le séjour. Les données sont ensuite anonymisées et envoyées aux organismes tutelles qui les utilisent principalement dans un objectif de financement des hôpitaux. Malgré l'anonymisation, une procédure de chaînage permet, grâce à une clé cryptée irréversible, de rattacher entre elles les hospitalisations d'un même patient dans différents établissements. Le chaînage offre donc la possibilité de reconstituer le parcours de soins hospitalier d'un patient à partir des données du PMSI. La nature des données du PMSI, leur exhaustivité et le mécanisme de chaînage en font donc une source d'information potentiellement très intéressante pour l'analyse des trajectoires de soins.

4.5 Conclusion

Dans le cadre des pathologies chroniques dont la prise en charge est fréquemment multidisciplinaire, la qualité et l'efficience des systèmes de soins reposent en partie sur la coordination des différents acteurs. Si les professionnels de santé peuvent s'appuyer à l'échelle individuelle sur des guides de bonne pratique, leur mise en œuvre dans un contexte coopératif peut se heurter à des problèmes organisationnels. Pour surmonter cette difficulté, il est nécessaire d'activer des mécanismes d'apprentissage des organisations. Cet apprentissage repose avant tout sur la connaissance du fonctionnement réel du système de soins. Plus précisément, les différents acteurs impliqués dans le processus de soins doivent disposer d'outils capables d'observer, décrire et classer le parcours des patients dans un espace de soins donné. À l'heure actuelle, peu de méthodes d'analyse des trajectoires de soins sont assez souple pour s'adapter à la diversité et la spécificité des objectifs poursuivis par l'apprentissage organisationnel. Pourtant, de nombreux systèmes d'information sont susceptibles de contenir les connaissances nécessaires à cette analyse. C'est le cas du PMSI en France qui collecte chaque année des données médico-administratives relatives à toutes les hospitalisations réalisées sur le territoire.

Nous proposons une approche générale d'extraction des connaissances reposant sur l'analyse formelle de concepts pour étudier les trajectoires de soins à partir de données médico-administratives. Dans le chapitre suivant, à partir de plusieurs exemples détaillés, nous illus-

trons l'intérêt de cette approche.

Chapitre 5

Analyse des trajectoires de soins par treillis de concepts

Sommaire

5.1 Introduction

La planification et l'organisation des soins jouent un rôle prépondérant dans l'amélioration de l'état de santé. À l'échelle des populations, même les thérapies de pointe ne suffisent pas si elle ne peuvent être mises en œuvre dans des conditions appropriées. De nombreux déterminants conditionnent l'efficacité d'une prise en charge : disponibilité de personnels entraîné, disponibilité des équipements, environnement et conditions de sécurité, coûts, proximité des structures de soins ... De surcroît, toutes ces contraintes doivent s'adapter aux besoins de la population et à ses caractéristiques économiques, démographiques et épidémiologiques.

C'est particulièrement vrai dans le domaine de la cancérologie. Le traitement du cancer repose en effet sur une prise en charge éminemment multidisciplinaire, dans laquelle interviennent de multiples professionnels de santé : oncologues, radiothérapeutes, chirurgiens, médecin traitant, infirmières, psychologues ... Ces professionnels exercent dans des lieux souvent différents, et leurs patients reçoivent des traitements longs, de haute technicité et coûteux. Bien que la

cancérologie délivre des soins planifiés obéissant à de nombreux protocoles, il n'est pas toujours évident que l'organisation du système de soins réponde de manière optimale à l'exigence de ces protocoles. Afin d'adapter l'organisation des soins, de favoriser la communication et la coopération entre équipes soignantes, afin d'augmenter le savoir organisationnel, les différents partenaires de santé publique doivent pouvoir s'appuyer sur des systèmes d'aide à la décision qui dressent un panorama stratégique des comportements intrinsèques du système de soins.

D'un coté, les systèmes de soins des pays développés peuvent être considérés comme « riches de données » car ils produisent des quantités considérables d'information par l'intermédiaire des dossiers électroniques, de la recherche clinique, des registres de pathologie, des systèmes d'information hospitaliers, etc. De l'autre coté, ils peuvent être vus comme « pauvres en connaissance » car ces données sont rarement intégrées dans de véritable système d'aide à la décision [Abidi, 2001]. En France, dans le domaine du cancer, plusieurs systèmes d'informations collectent des informations potentiellement utiles comme le PMSI où les registres des cancers. Néanmoins, ces systèmes souffrent de plusieurs limites. Soit ils ne sont pas adaptés à la surveillance des pathologies au long cours comme c'est le cas pour le PMSI, qui est centré sur les hospitalisations et non pas sur le patient. Soit ils ne couvrent que certaines pathologies et des zones géographiques restreintes, comme les registres, et ne peuvent être raisonnablement déployés à une échelle nationale.

Dans ce chapitre, nous mettons en œuvre plusieurs approches basées sur l'AFC pour explorer, à partir des données du PMSI, les trajectoires de soins dans le domaine de la cancérologie. Tout d'abord, nous considérons le système de soins comme une organisation virtuelle dont les membres (des hôpitaux) fonctionnent en réseau, partageant la prise en charge de patients communs. La section 5.2 présente la problématique des réseaux sociaux et les contributions de l'AFC dans ce domaine. Dans la section 5.3 nous montrons les apports de l'AFC pour découvrir et visualiser l'organisation virtuelle du système de soins. La section 5.4 s'inscrit dans la continuité de cette approche et met l'accent sur l'utilisation conjointe de la stabilité et du support des concepts pour améliorer la détection des filières de soins. La section 5.5 décrit la construction d'une classification des profils de soins à l'aide de treillis de concepts. Dans cette approche, nous nous intéressons à la construction non supervisée d'un SCM à partir de données relatives à des diagnostics établis lors d'hospitalisations successives de patients atteints de cancer du sein. Enfin, dans la dernière section, nous proposons une approche d'analyse séquentielle des trajectoires de soins par treillis de concepts.

5.2 Analyse de réseaux sociaux par AFC

Un réseau social est un ensemble d'entités sociales telles que des personnes ou des organisations sociales reliées entre elles par des liens créés lors des interactions sociales. L'analyse de réseaux sociaux conçoit les relations sociales en termes de nœuds et de liens et s'appuie sur la théorie des graphes. Typiquement, un réseau social est représenté par une matrice d'adjacence binaire dans laquelle chaque ligne et chaque colonne représentent un membre du réseau. L'analyse des réseaux sociaux couvre un ensemble de problèmes dont les principaux sont : l'identification des communautés, l'étude de la dynamique d'un réseau, l'identification des rôles des nœuds et des communautés qui forment un réseau social, l'étude et la prédiction de l'évolution des réseaux. Le succès de l'analyse des réseaux sociaux repose en grande partie sur la visualisation. Les images fournies par la théorie des graphes attirent l'attention sur les propriétés structurelles des réseaux qui pourraient passer inaperçues autrement. La figure 5.1 montre un exemple de graphe de réseau social.

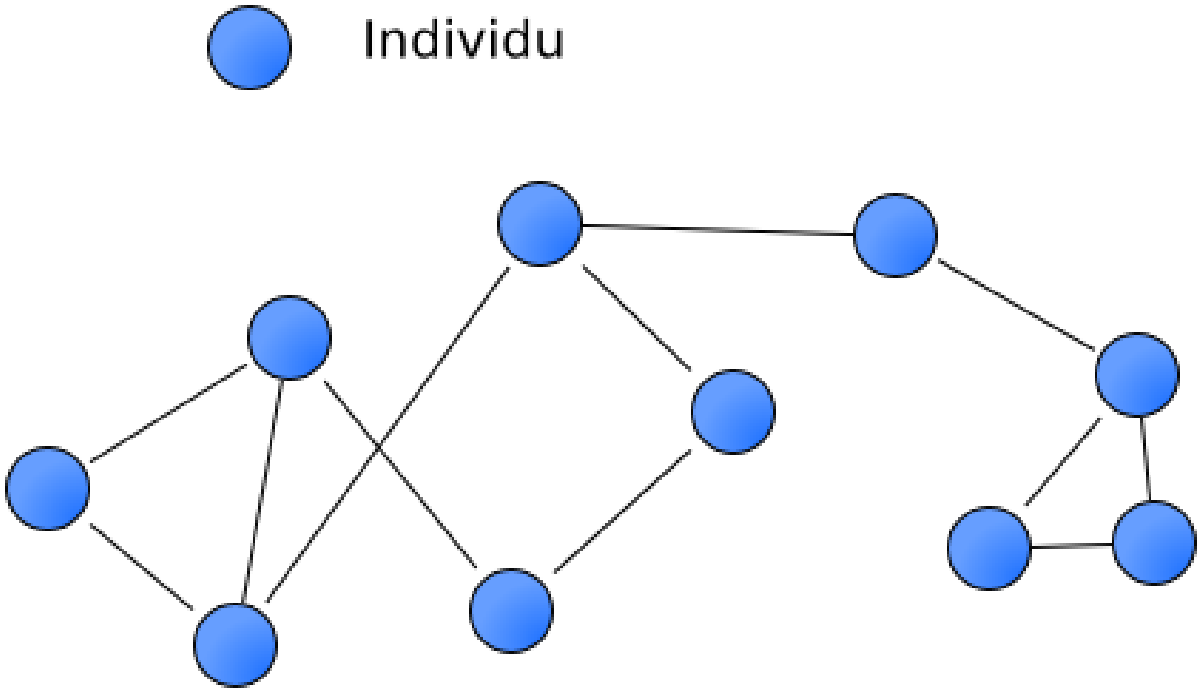

FIG. 5.1 – Un exemple de graphe de réseau social : les nœuds sont des individus et les arcs matérialisent un lien entre deux individus (par exemple d'amitié).

Pour autant, ce modèle de représentation présente certaines limites. En ne tenant compte que de liens qui connectent des paires d'acteurs, il restreint les possibilités de dévoiler des formes importantes de liens social [Freeman and White, 1993]. En particulier il ne permet pas

de distinguer les relations sociales liant des membres deux à deux de celles qui englobent un plus grand nombre de participants. Considérons par exemple trois amis d'enfance : Pierre, Paul et Jacques. Au collège, Pierre était très proche de Paul, tandis qu'au lycée, Paul et Jacques ne se quittaient jamais. On a là trois paires d'individus qui ont été liés en différents temps et différents lieux par une relation d'amitié. A l'opposé, supposons que Sandrine, Émilie et Camille étaient inséparables à la faculté lorsqu'elles formaient un petit groupe de rock. Il s'agit dans ce cas d'une communauté qui est une forme de relation sociale dépassant la simple juxtaposition de paires de liens disjoints. Alors que la théorie des graphes est tout à fait adaptée à la visualisation des liens qui unissent Pierre, Paul et Jacques, elle ne permet pas de représenter simplement l'essence de la communauté formée par Sandrine, Émilie et Camille.

Parmi les solutions envisagées pour dépasser ces limites, Freeman *et al.* ont proposé d'utiliser l'analyse formelle de concept [Freeman and White, 1993]. En modélisant un ensemble d'individus et des activités sociales auxquels ils ont participé par un contexte formel (les individus comme objets, les activités comme attributs), ils visualisent et interprètent ce réseau social à l'aide du treillis de concepts associé. Au delà de la visualisation, ce dernier permet aussi de révéler et hiérarchiser les rôles joués par les membres du réseau. Cependant, Freeman *et al.* notent que les gains apportés par la visualisation diminuent rapidement lorsque le contexte formel et le treillis atteignent une certaine taille.

Les travaux menés par Duquenne [Duquenne, 1996b, Duquenne, 1996a] préfigurent ceux de Stumme [Stumme et al., 2002] sur les treillis icebergs. Duquenne propose une simplification et une approximation du treillis de concept basées sur la cardinalité des extensions – en fait le support– et sur une saturation du treillis à partir de « quasi-implications », qui sont des règles d'association à confiance élevée. Par la suite, Stumme *et al.* [Stumme et al., 2002] ont introduit les icebergs de concepts, ainsi que l'algorithme `TITANIC` qui permet de les calculer efficacement. Les icebergs résolvent en partie les problèmes de visualisation de grands treillis en élaguant les concepts de faible support. Cette approche repose sur l'assomption que les connaissances intéressantes sont représentées par des concepts fréquents. Cette hypothèse n'est pas toujours vérifiée, notamment en matière de réseaux sociaux.

En s'intéressant aux communautés épistémologiques, Roth, Kuznetsov *et al.* reprennent la notion de stabilité pour découvrir et représenter de petites communautés d'acteurs caractérisées par une grande cohésion [Roth et al., 2006, Roth, 2006, Kuznetsov et al., 2007]. Les communautés épistémologiques sont des groupes d'agents possédant un intérêt commun pour un ou des domaines d'étude particuliers. Par exemple, certains chercheurs sont mathématiciens et s'intéressent à la théorie des ensembles ordonnés. D'autres sont informaticiens et se préoccupent plus spécialement de fouille de données. Enfin, certains chercheurs spécialisés en AFC sont concernés par ces deux domaines. Ces champs d'investigation, associés à leurs acteurs,

sont des communautés épistémologiques. Or, les concepts formels se prêtent particulièrement bien à la représentation de telles notions si on modélise les acteurs comme objets et leurs domaines de recherche comme attributs d'un contexte formel. De plus, les treillis de concepts permettent de dévoiler la structure des communautés épistémologiques. Pour Roth *et al.*, une communauté épistémologique est d'autant plus réelle qu'elle ne dépend pas de quelques agents en particulier : une communauté « réelle » doit être stable malgré des informations bruitées. En dehors de la résistance au bruit, une communauté ne se décompose pas en sous-groupes dès lors que quelques uns de ses membres la quittent. Enfin, il existe des communautés formées d'un petit nombre d'agents très soudés. La représentation par treillis des communautés épistémologiques devient problématique si l'on a à faire à des contextes de taille importante. Les icebergs permettent de les visualiser graduellement en fonction de leur taille, mais ne tiennent pas compte de leur stabilité. C'est pourquoi Roth et Kuznetsov proposent un élagage de treillis par la stabilité [Roth et al., 2006, Kuznetsov et al., 2007].

5.3 Visualisation de l'organisation virtuelle du système de soins

Le traitement des maladies chroniques est souvent multidisciplinaire. Il fait intervenir plusieurs acteurs du champ de la santé, aux spécialités différentes : médecins spécialistes, généralistes, infirmières, kinésithérapeutes... Le parcours des patients chroniques passe également par plusieurs institutions, surtout lorsque leur traitement nécessite le recours à des techniques ou équipements spécialisés disponibles uniquement dans certains centres. C'est par exemple le cas en cancérologie, où la prise en charge est une séquence d'actes diagnostiques (scanner, IRM, endoscopie), chirurgicaux (ablation de la tumeur), médicaux (radiothérapie, chimiothérapie). Après une première phase de traitement, les patients peuvent entrer à nouveau en contact avec le système de soins, soit pour des examens de surveillance, soit lors de complications dues à la maladie elle-même ou aux traitements. Cette prise en charge s'étale fréquemment sur plusieurs mois, voire plusieurs années. Pour assurer une qualité des soins optimale, tout en gérant au mieux les ressources disponibles, il est indispensable que cette prise en charge soit coordonnée. Pour répondre à la demande de soins, les professionnels et les institutions s'organisent en réseaux. Ces réseaux peuvent être formalisés. C'est d'ailleurs une obligation formulée par le plan cancer lancé en 2003 en France. Ainsi, les établissement de soins impliqués dans le traitement du cancer en Lorraine se sont de longue date organisés au sein du réseau Oncolor. Malgré tout, le parcours du patient dans le système de soins n'est pas aussi déterministe que peuvent le laisser penser les recommandations ou la planification théorique de l'offre de soins. Cet état de faits dépend d'un certain nombre de facteurs, à commencer par l'incertitude de l'évolution de la maladie, mais aussi à des contraintes structurelles, économiques, démographiques et sociales.

Ainsi, le fonctionnement du système de soins repose beaucoup sur une auto-organisation virtuelle, non formelle, des prestataires de soins. Pour pouvoir renforcer la cohérence du système et optimiser la coordination des acteurs, ces derniers doivent pouvoir s'appuyer sur des outils d'aide à la décision qui décrivent le comportement intrinsèque de ce réseau « virtuel » de soins.

5.3.1 Topologie du réseau de soins

Nous nous sommes intéressés à l'organisation du système de soins hospitalier en Lorraine dans le domaine de la cancérologie. A partir des données du PMSI, nous avons construit un contexte formel dans lequel les objets sont des patients atteints de cancer, et les attributs des hôpitaux qui les ont traités pour leur cancer. Une étape de pré-traitement a consisté à sélectionner tous les séjours des patients résidant en Lorraine correspondant au traitement d'un cancer au cours de l'année 2003. Cette sélection a été réalisée sur la base des diagnostics principaux, des diagnostics reliés et du code postal de résidence. Le contexte obtenu contient 31578 patients et 298 hôpitaux. Un treillis iceberg a ensuite été calculé en prenant pour seuil de support correspondant à 50 patients. Pour ce faire, nous avons utilisé l'algorithme `TITANIC` implémenté dans Galicia[Valtchev et al., 2003]. L'iceberg obtenu contient 86 concepts. Il est représenté sur la figure 5.2. Pour plus de clarté, le concept $\perp$ n'est pas affiché. Les étiquettes des concepts font apparaître les noms des hôpitaux de l'intension et leur support absolu, qui représente le nombre de patients communs aux hôpitaux de l'intension.

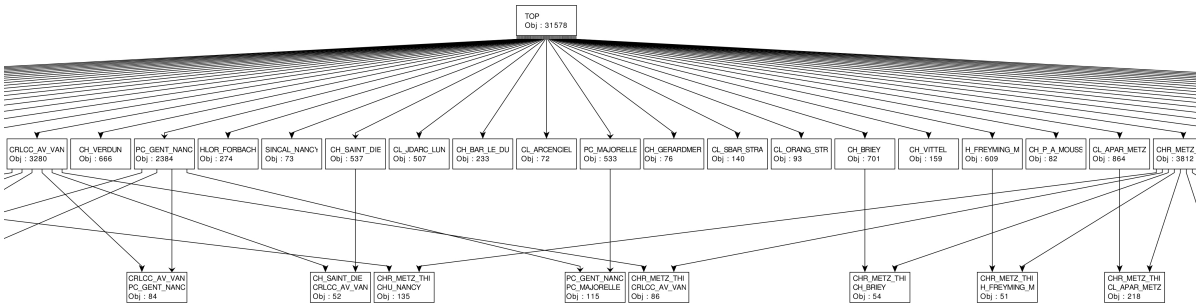

FIG. 5.2 – Iceberg montrant l'organisation virtuelle du réseau hospitalier Lorrain en cancérologie.

Un premier commentaire peut être fait sur la forme générale du treillis : il est plus large que haut. En effet le contexte formel est épars et les attributs sont peu corrélés. Cela signifie que les filières de soins sont compartimentées et que les patients sont rarement hospitalisés dans plus de deux hôpitaux différents. Le concept $\top$ (top) fait apparaître les 31578 patients lorrains hospitalisés pour cancer en 2003. Son intension est vide, comme on peut s'y attendre, aucun hôpital n'a accueilli tous ces patients. Les descendants immédiats de $\top$ ont tous une

intension de taille 1 : aucun hôpital ne partage tous ses patients avec un autre, ou, si c'est le cas, le nombre de patients est inférieur à 50. Les concepts de rang deux sont des atomes, c'est-à-dire des ascendants immédiats de $\perp$. Leur intension est toujours une paire d'hôpitaux. Leur support donne une idée de la force de la coopération entre deux hôpitaux. Plus il est élevé, plus ces hôpitaux partagent de patients et plus on peut penser qu'ils exercent une forme réelle de collaboration. En outre, il n'y pas de collaboration impliquant plus de deux hôpitaux et plus de 50 patients. La figure 5.3 montre un extrait de l'iceberg de la figure 5.2. Le CHR de Metz-Thionville (CHR_METZ_THI) a traité 3812 patients. Cet hôpital a de nombreuses coopérations : le concept CHR_METZ_THI compte quatre sous-concepts. Ainsi, le CHR de Metz-Thionville partage 117 patients avec l'hôpital Belle-Isle de Metz (H_BELL_METZ). L'intensité de la coopération Belle-Isle – CHR de Metz semble plus forte que celle de la coopération du CHR de Metz avec l'hôpital de Briey (CH_BRIEY) : Belle-Isle partage 24%(117/481) de ses patients avec le CHR, Briey 8%(54/701) seulement. Le treillis montre bien la situation particulière du CHR de Metz. Ce gros établissement dispose d'équipes et d'un plateau technique spécialisés dans le traitement du cancer et joue un rôle de recours dans son bassin géographique. D'autres établissements, comme Vittel (CH_VITTEL), n'affichent pas de coopération. C'est souvent parce qu'ils ne traitent qu'un petit nombre de patients et que le seuil choisi pour construire l'iceberg ne permet pas de visualiser d'éventuelles interactions. Parfois, cependant, cela correspond à des établissements capables de travailler dans une relative autonomie car ils sont les seuls, dans un bassin géographique isolé, à réunir les ressources nécessaires au traitement du cancer.

L'iceberg peut être redessiné en supprimant les concepts $\top$ et $\perp$ comme sur la figure 5.4. Les concepts les plus sombres représentent les co-atomes du treillis, c'est-à-dire la couverture inférieure de $\top$. Les plus clairs correspondent aux coopérations entre deux établissements. L'orientation des arcs indique la relation superconcept-sousconcept. Les établissements n'ayant pas de coopérations ne figurent pas sur le schéma. Ce mode de représentation fait clairement apparaître la topologie de l'auto-organisation des établissements lorrains en un réseau virtuel. Tout d'abord, trois établissements occupent une place centrale dans le réseau : le CHU de Nancy, le CHR de Metz et le Centre de Lutte Contre le Cancer Alexis Vautrin (CAV) représentés respectivement par les concepts CHU_NANCY, CHR_METZ_THI et CLCC_AV_VAN. Le CHU de Nancy est le plus gros établissement de la région Lorraine. C'est un hôpital universitaire accueillant de nombreuses équipes spécialisées dans le traitement du cancer. Il dispose d'équipements lourds (imagerie, plateaux chirurgicaux, chambres stériles ...). C'est également un centre de recherche qui étudie de nouveaux traitements. Le CHU de Nancy est un centre de référence dans la région, principalement pour la partie sud. Bien que non universitaire, le CHR de Metz-Thionville a un profil similaire et a un rôle de recours pour le nord de la Lorraine.

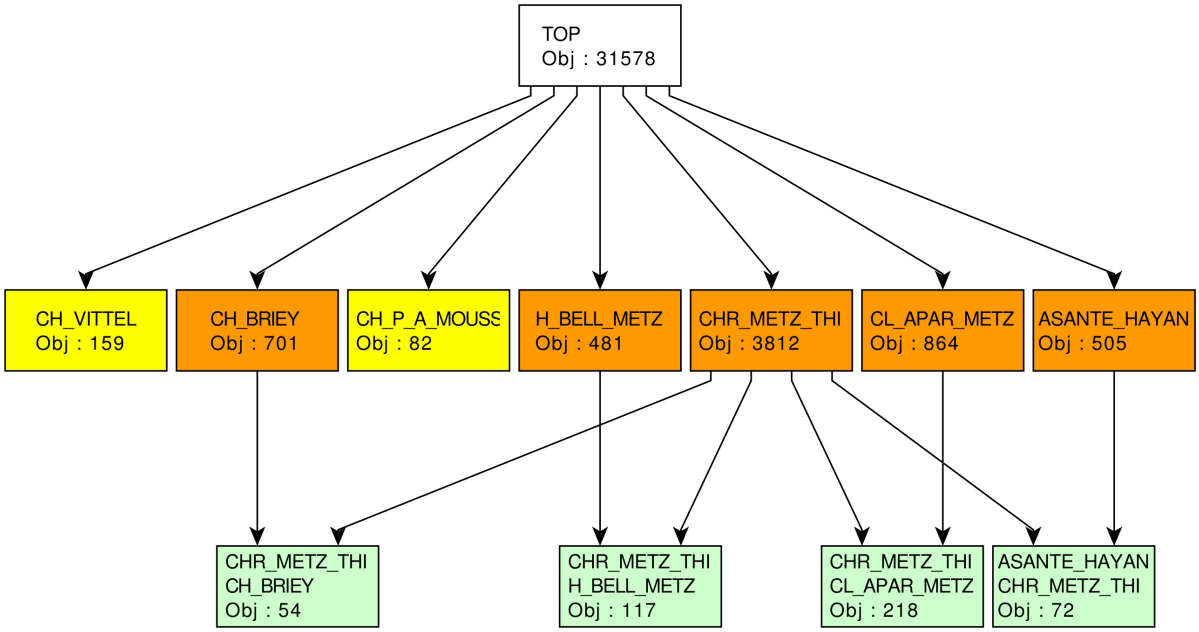

FIG. 5.3 – Un extrait de l'iceberg de la figure 5.2.

Enfin, le CAV est un centre spécialisé dans la lutte contre le cancer doté de personnels entraînés et d'équipement de pointe dans ce domaine. Ces trois hôpitaux sont ceux qui traitent le plus grand nombre de patients dans la région et leurs concepts ont les supports les plus élevés. Ils entretiennent tous les trois de nombreuses collaborations, essentiellement avec leurs voisins géographiques.

Les lignes en pointillés illustrent les contraintes qui façonnent la topologie du réseau. Le treillis peut être découpé en cinq composantes connexes (1,2,2',3 et 4) qui constituent autant de « sous-réseaux » locaux. La composante 1 rassemble trois établissements situés dans le département des Vosges, au sud-est de la Lorraine. Dans une certaine mesure, l'hôpital d'Épinal (CH_EPINAL) y joue un rôle de référence. C'est le seul établissement du département à être équipé d'appareils de radiothérapie. Il entretient des collaborations avec deux autres hôpitaux des Vôsges, mais aussi avec le CHU de Nancy et le CAV, également situé à Nancy. Les composantes 2 et 2' regroupent les concepts représentant des établissements qui sont presque tous situés à Nancy. C'est là qu'on retrouve la coopération impliquant le plus grand nombre de patients : 632 entre le CHU de Nancy et le CAV. Cela s'explique par le fait que le CHU de Nancy n'est pas équipé d'appareil de radiothérapie, et qu'il adresse la plupart des patients devant en bénéficier au CAV. La composante 2' fait apparaître que trois cliniques du secteur privé de Nancy, tout en collaborant avec le secteur public, constituent un sous-ensemble d'interactions privilégiées. La composante 3 montre comment les hôpitaux du Nord de la Lorraine

s'organisent en « étoile » autour du CHR de Metz-Thionville. Enfin, la composante 4 traduit le fait qu'un nombre conséquent de patients frontaliers de l'Alsace sont traités par deux établissements de Strasbourg : le CHU de Strasbourg (CHU_STRAS) et le centre de lutte contre le cancer Paul Straus(CLCC_PS_STR). Elle montre également que l'hôpital de Sarreguemines (CH_SARREGUEM), en Lorraine, privilégie ses contacts avec Strasbourg. On peut finalement souligner la séparation entre les établissements du Sud lorrain (composantes 1,2 et 2') et ceux du Nord (composantes 3 et 4). Les seuls échanges existant s'opèrent entre les trois établissements les plus gros et les plus spécialisés de Lorraine : CHU de Nancy, CAV et CHR de Metz-Thionville. Ils impliquent d'ailleurs assez peu de patients en regard des effectifs traités par chacun d'entre eux.

La visualisation de l'organisation virtuelle du réseau de soins par AFC est un support riche d'informations pour l'analyste. L'iceberg dessine la topologie du réseau et fournit des données quantitatives permettant de mesurer des flux de patients. En cela, il formalise la dynamique des interactions entre établissements et rend explicite une réalité dont les acteurs du système de soins n'ont souvent qu'une connaissance partielle. L'analyste, en mobilisant ses connaissances du domaine, est à même d'interpréter ces schémas sans être très familier de l'AFC. Il peut expliquer les facteurs qui déterminent la forme générale du treillis et qui sont à l'origine des interactions entre hôpitaux : géographie, équipements, spécialisation des équipes ...

5.3.2 Raffinement de l'analyse

En s'inscrivant dans le cycle d'extraction des connaissances, il est possible d'affiner l'analyse en modifiant le contexte formel initial. On peut par exemple s'intéresser à une sous population de patients traités pour des localisations tumorales spécifiques. Pour ce faire, nous avons enrichi le contexte formel d'attributs qui, pour chaque patient, indiquent la (ou les) localisation(s) tumorale(s) par appareil anatomique (poumon, sein, système digestif ...). Un nouvel iceberg a été construit cette fois avec un support minimum correspondant à 10 patients afin de pouvoir explorer précisément l'organisation du réseau par localisation tumorale. L'analyste peut sélectionner une partie du treillis correspondant à une localisation précise : il suffit pour cela d'extraire l'idéal du plus grand concept contenant cette localisation dans son intension. La figure 5.5 montre l'idéal du concept SEIN dont l'extension correspond à la population des patients atteints de cancer du sein. Le schéma s'interprète de la même manière que la figure 5.4. On remarque que la topologie du réseau est cette fois différente. En particulier, le CHU de Nancy y joue un rôle tout à fait marginal. En revanche, le CAV (concept CRLCC_AV_VAN) occupe une place centrale dans le réseau. Il est l'établissement qui accueille le plus grand nombre de patients et interagit avec sept autres institutions. Le CAV est en fait une référence régionale dans le traitement du cancer du sein. On peut également noter que l'hôpital de Sarreguemines est ici

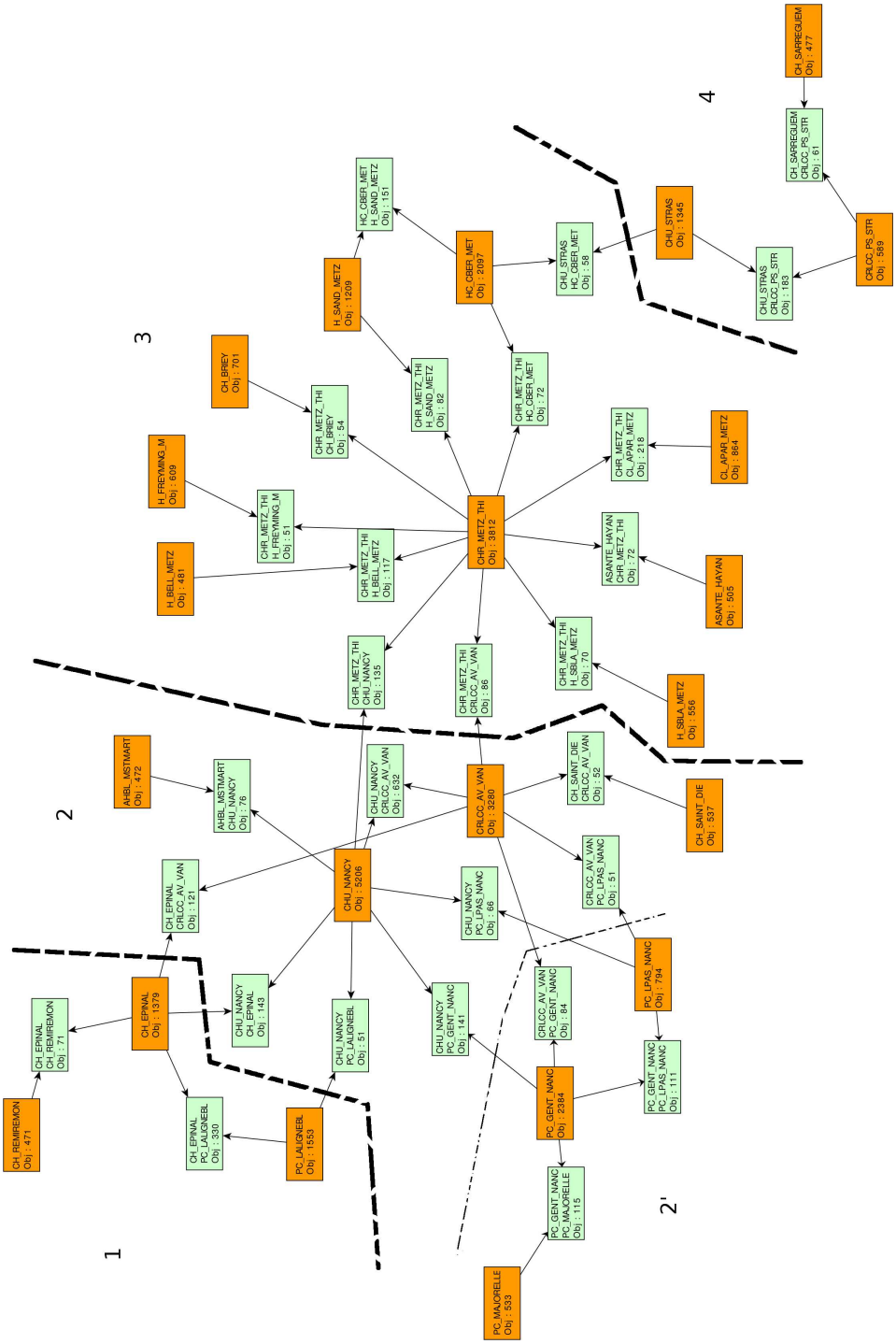

FIG. 5.4 – Topologie virtuelle du réseau de soins hospitalier lorrain pour le traitement du cancer.

complètement déconnecté du reste du réseau et travaille exclusivement avec un établissement alsacien (à droite sur la figure).

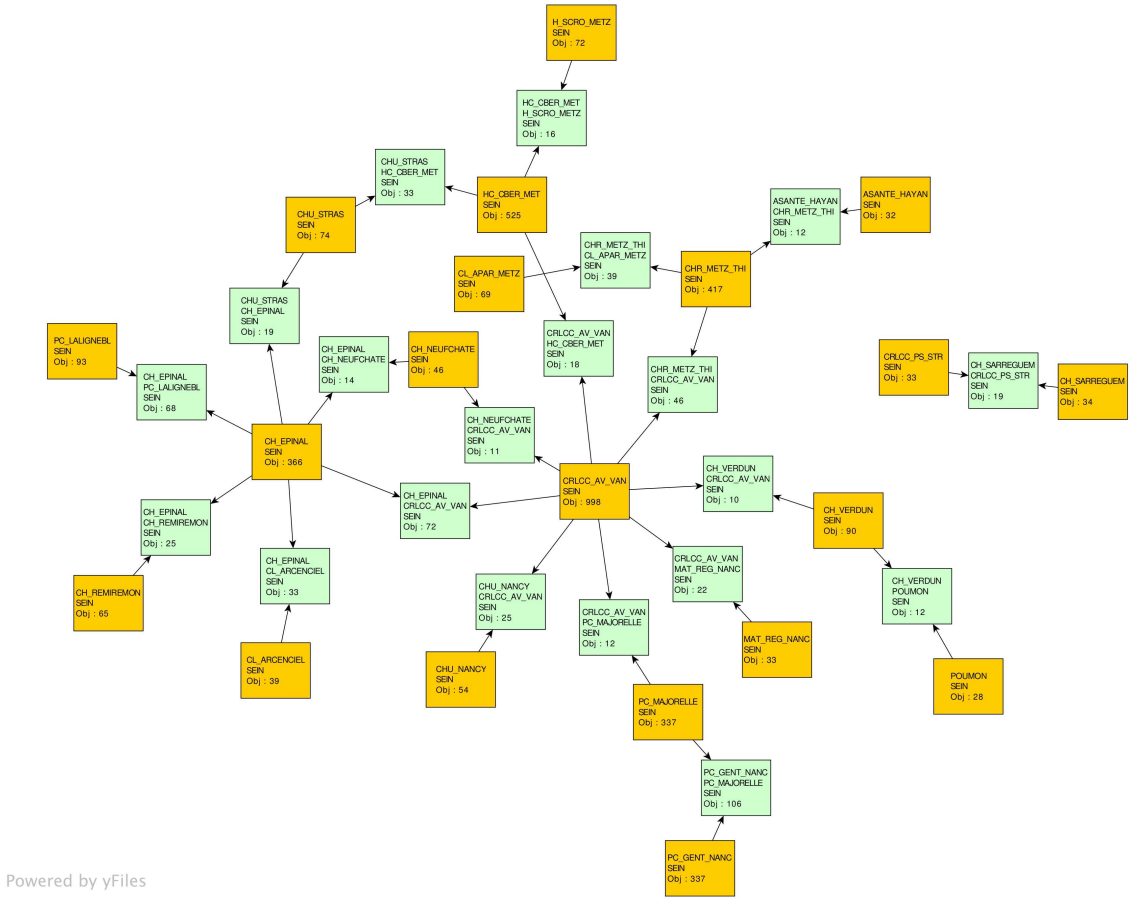

FIG. 5.5 – Organisation virtuelle des hôpitaux lorrains pour le traitement du cancer du sein.

5.3.3 Interrogation du treillis

L'enrichissement du contexte formel augmente le nombre de concepts du treillis. Parallèlement, il faut diminuer le support minimum pour accéder à une information intéressante et la visualisation intégrale de l'iceberg perd alors en lisibilité : en abaissant le support minimum de 50 à 10 patients, on passe de 86 à 630 concepts. Néanmoins il est toujours possible, comme la section 5.3.2 le suggère, de ne visualiser qu'une partie de l'iceberg qui peut servir de support d'exploration. L'analyste peut par exemple être intéressé par la nature de la coopération entre deux établissements. La figure 5.6 montre la partie du treillis qui correspond à la requête « quelle est la nature de la coopération entre le CHU de Nancy et le CHR de Metz-Thionville ? ». Il

s'agit du filtre et de l'idéal du suprémum des concepts dont l'intension contient CHU_NANCY et CHR_METZ_THI. Pour plus de clarté, le schéma utilise un étiquetage réduit des attributs. On y lit que 135 patients ont été hospitalisés à la fois au CHU de Nancy et au CHR de Metz-Thionville. Parmi eux, 19 ont une tumeur du sang (BLOOD), 10 une tumeur du poumon (LUNG), 41 une tumeur cérébrale (BRAIN), 24 une tumeur digestive et 20 une tumeur de l'appareil urinaire (UADT). On remarque également que 10 de ces patients ont été hospitalisés dans un troisième établissement (CRLCC_AV_VAN).

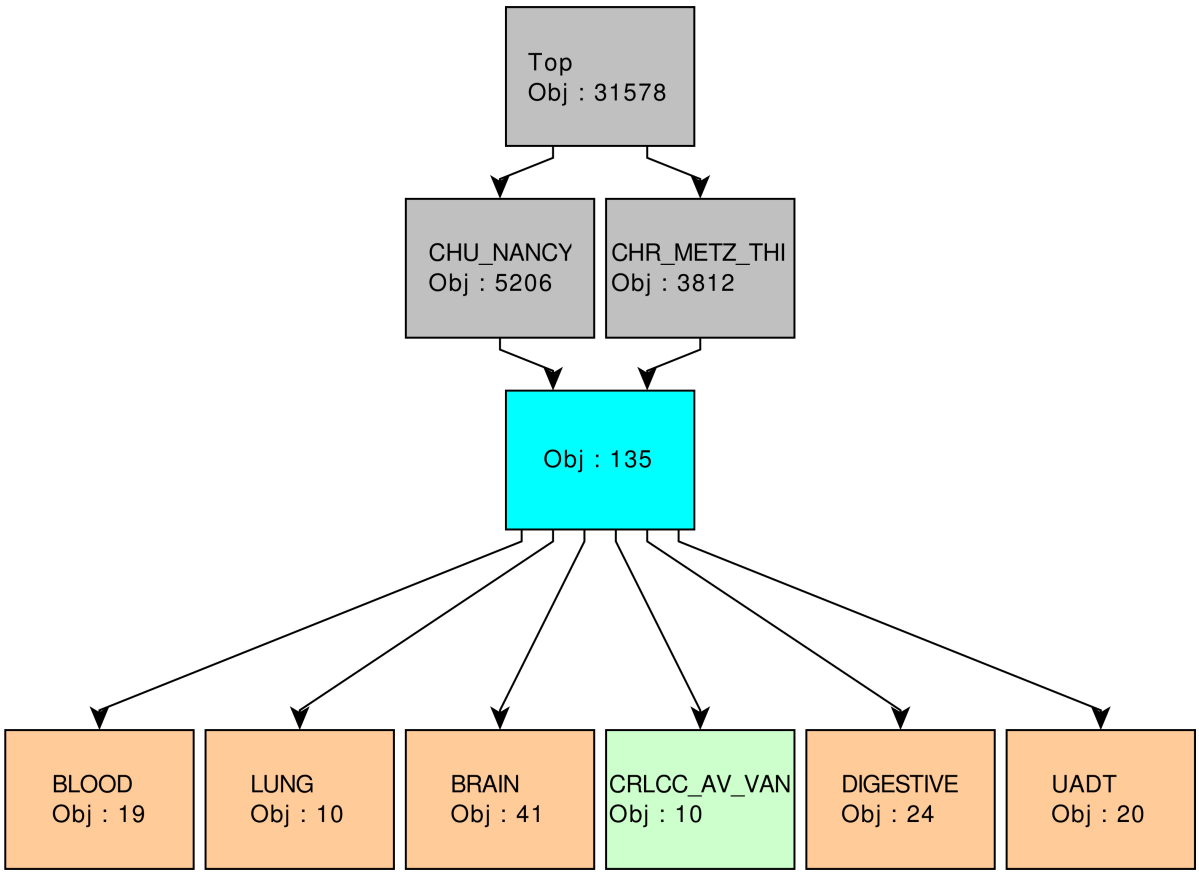

FIG. 5.6 – Détail de la coopération entre le CHU de Nancy et le CHR de Metz-Thionville.

Formellement on peut définir le processus d'interrogation du treillis comme suit. Soit $\mathbb{K} = (\mathtt{G}, \mathtt{M}, \mathtt{I})$ un contexte formel, $\underline{\mathfrak{B}}(\mathbb{K})$ son treillis de concepts, et `minsupp` un seuil de support minimum. L'analyste définit une requête comme un ensemble d'attributs $\mathtt{R} \subseteq \mathtt{M}$. On appelle concept requête le concept $\mathtt{C_R} = \bigvee\{(\mathtt{X}, \mathtt{X}') \in \underline{\mathfrak{B}}(\mathbb{K}) | \mathtt{R} \subseteq \mathtt{X}'\}$. Le résultat de la requête est l'intersection de l'iceberg défini par `minsupp` avec la réunion du filtre et de l'idéal de $\mathtt{C_R}$, soit :

$$\{\mathtt{C} \in \downarrow \mathtt{C_R} \cup \uparrow \mathtt{C_R} | \sigma(\mathtt{C}) \geq \mathtt{minsupp}\}$$

.

5.4 Stabilité des trajectoires de soins

5.4.1 Difficulté de choisir un seuil de support

Le choix du seuil de support influence fortement les connaissances qui vont être découvertes à la lecture de l'iceberg. Il doit être suffisamment bas pour véhiculer des informations intéressantes tout en étant suffisamment élevé pour que l'analyste puisse les traiter. Dans cette section, nous prenons en exemple un treillis construit à partir d'un contexte où les objets sont des patients lorrains atteints de cancer en 2005 et les attributs des hôpitaux qui les ont accueillis. Le treillis complet compte 865 concepts. La figure 5.7 montre la distribution cumulée du support des concepts du treillis. Près de 90% des concepts ont un support inférieur à 10.

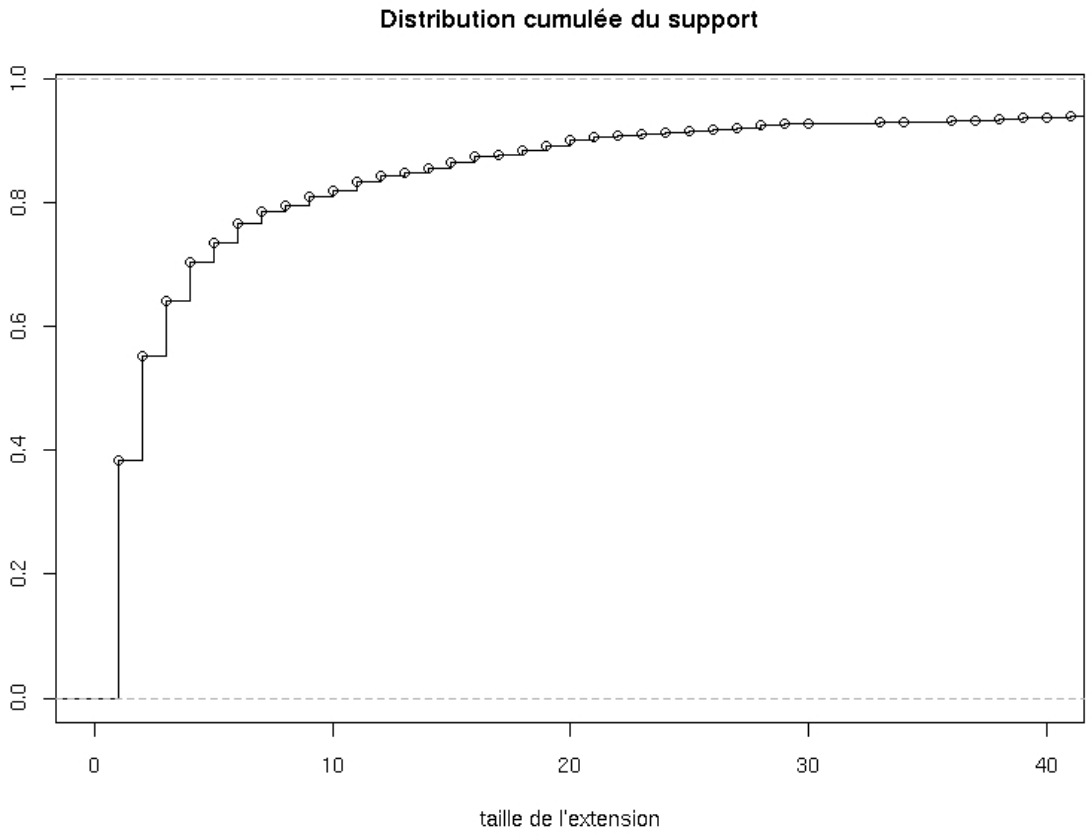

FIG. 5.7 – Distribution cumulée du support des concepts.

Le tableau 5.1 montre la distribution du support (en valeur absolue) des concepts en fonction

Nombre de patients \ Taille de l'intension	0	1	2	3	≥ 4
< 10	0	40	285	297	76
$[10, 20[$	0	13	52	7	0
$[20, 30[$	0	10	20	0	0
$[30, 40[$	0	6	2	0	0
> 40	1	34	20	0	0

TAB. 5.1 – Répartition du nombre de concepts par taille de l'extension et de l'intension.

de la taille de leur intension.

Si l'on choisi un seuil correspondant à 20 patients, on obtient déjà 93 concepts, ce qui commence à poser des problèmes de visualisation. On risque par ailleurs de passer à coté de connaissances intéressantes :

- 13 hôpitaux accueillent au moins 10 patients.
- 52 coopérations entre deux hôpitaux impliquent au moins 10 patients,
- 7 concepts représentent des interactions entre trois hôpitaux pour au moins 10 patients.

Ces concepts sont potentiellement intéressants dans notre étude car la réglementation française récente soumet l'autorisation d'exercice de la cancérologie à des seuils minimaux d'activité. Or ces seuils sont relativement bas (entre 20 et 30 interventions par an). Dans l'hypothèse d'une réorganisation de l'offre de soins, il est indispensable de connaître les interactions entre établissements, même si elles concernent peu de patients, afin de les renforcer le cas échéant.

Ici, le support a un faible pouvoir discriminant pour évaluer l'intérêt de concepts « peu fréquents ». Certains, qui n'apparaîtraient pas dans l'iceberg, sont potentiellement porteurs de connaissances singulières :

- ils traduisent l'interaction d'un établissement partageant presque tous ses patients avec un autre,
- ils concernent des hôpitaux qui traitent peu de patients et qui n'ont pas de d'interactions avec d'autres,
- ils illustrent des interactions entre trois établissements.

Diminuer le seuil de support pour construire l'iceberg permettrait de faire apparaître ce type de concepts, aux dépends de la lisibilité. Mais, au-delà, le support à lui seul n'est pas capable de distinguer des concepts ayant ces caractéristiques pour au moins deux raisons :

- on ne dispose que de 10 valeurs distinctes pour discriminer les 72 concepts dont le support se situe entre 10 et 19 patients,
- le support est une mesure de fréquence qui ne véhicule aucune information sur la structure du treillis, notamment sur les liens entre un concept et ses sous-concepts.

5.4.2 Analyse de la stabilité

Dans cette section, nous analysons la stabilité des concepts de l'ensemble du treillis. La figure 5.8 montre la distribution cumulée de la stabilité.

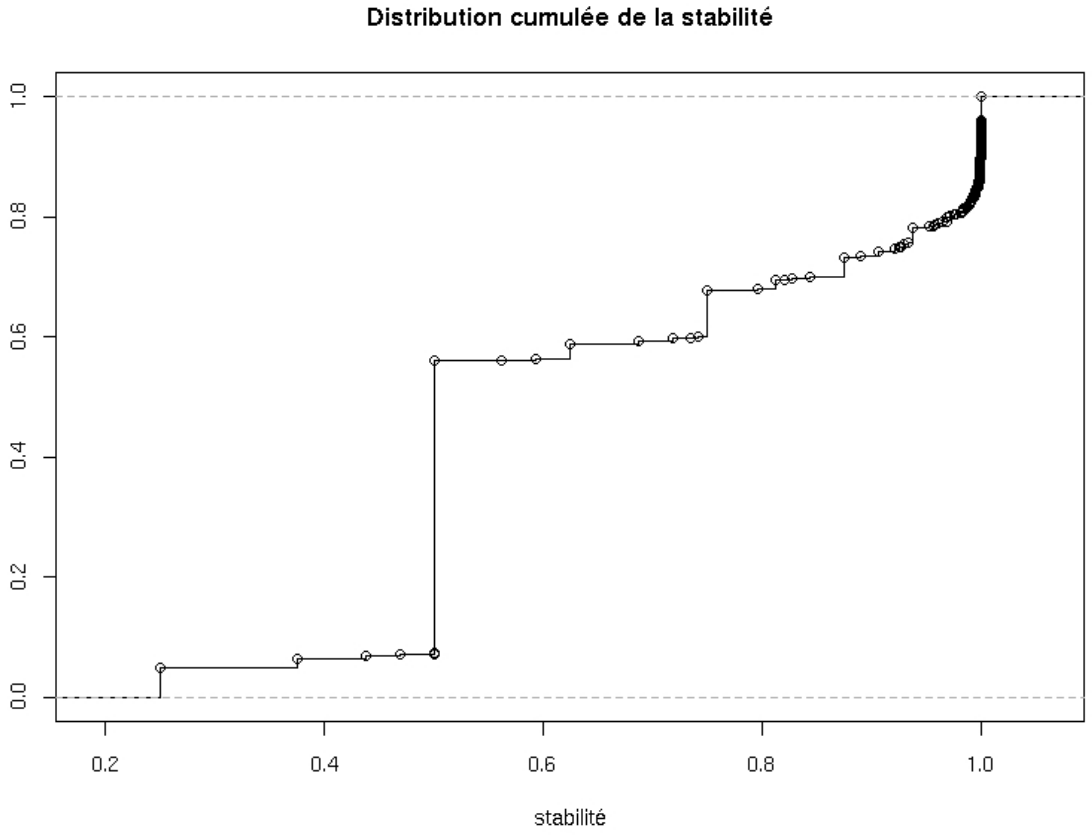

FIG. 5.8 – Distribution cumulée de la stabilité des concepts du treillis.

La pente de cette courbe apporte un certain nombre d'éléments. D'abord, quelques concepts (10%) ont une stabilité très basse (<0.5). Leur support est généralement faible (entre 2 et 6 patients) et leur idéal comporte entre 4 et 5 concepts (dont $\perp$ et eux-mêmes). Ensuite, un grand nombre de concepts (environ 50%) ont une stabilité de 0.5. Le concept $\perp$ ayant une extension vide, il s'agit là des concepts dont l'extension ne contient qu'un seul patient. La courbe est alors de nouveau légèrement croissante jusqu'à une dernière rupture qui correspond à des concepts ayant une stabilité très proche de 1. Ces derniers sont les concepts dont le support est le plus grand. En réalité, le $90^{\text{ème}}$ percentile de la distribution de la stabilité vaut 0.9999653.

La figure 5.9 est un diagramme de dispersion de la stabilité et du support des concepts.

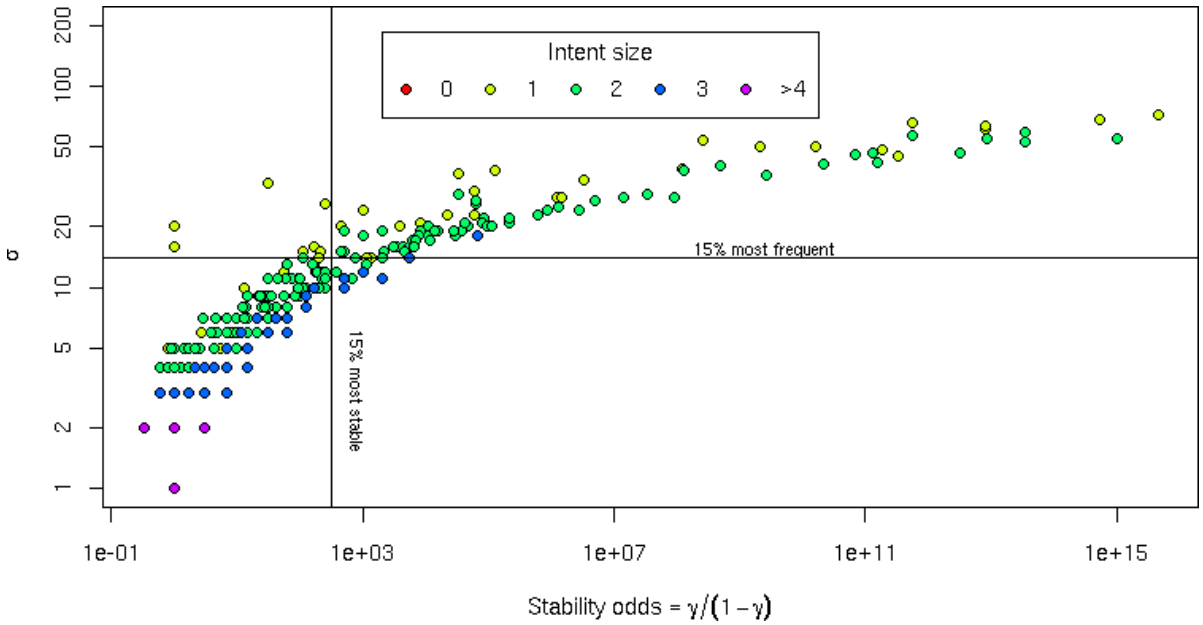

FIG. 5.9 – Diagramme de dispersion de la stabilité et du support.

Les valeurs sont présentées sur une échelle logarithmique pour mieux rendre compte de leur dispersion. De plus, nous préférons utiliser une mesure dérivée de la stabilité qui permet de mieux visualiser les valeurs proche de 1 : la cote de stabilité,

Définition 42. *Soit un concept* $\mathtt{C} \neq \bot$*, on appelle cote de stabilité de* $\mathtt{C}$ *le rapport :*

$$o_\gamma(\mathtt{C}) = \frac{\gamma(\mathtt{C})}{1 - \gamma(\mathtt{C})} \tag{5.1}$$

La cote de stabilité est le rapport du nombre de sous-ensembles de l'extension d'un concept qui appartiennent à sa classe d'équivalence sur le nombre de ceux qui n'y appartiennent pas. La stabilité a potentiellement un meilleur pouvoir discriminant que le support, surtout pour les concepts rares. On compte 57 valeurs distinctes de stabilité pour concepts dont le support absolu est entre 10 et 19. La figure 5.9 montre également les $85^{\text{èmes}}$ percentiles du support et de la cote de stabilité, ce qui permet de distinguer deux types de concepts :

- les concepts instables et fréquents,
- les concepts stables et rares.

5.4.3 Concepts instables et fréquents

Le quadrant supérieur gauche de la figure 5.9 fait apparaître 7 concepts que l'on peut considérer comme fréquents et instables. Leur intension est un singleton et ils représentent des

hôpitaux de deuxième ou troisième ligne en terme de recours pour le traitement du cancer. Cette notion de recours renvoie au degré de technicité et de spécialisation des soins qui y sont prodigués. Les hôpitaux de première ligne traitent le tout venant et adressent leurs patients vers des hôpitaux de deuxième ou troisième niveaux de recours lorsque des soins plus spécialisés ou nécessitant un plateau technique plus lourd sont nécessaires. Ces 7 hôpitaux partagent la plupart de leurs patients avec d'autres établissements, donnant naissance à deux types de coopération :

- exclusive,
- compartimentée.

La figure 5.10 montre les idéaux de deux concepts fréquents instables. Les deux hôpitaux concernés sont des hôpitaux de deuxième ligne : la Clinique de l'Arc-en-ciel (CL_ARCENCIEL) à gauche, la Maternité Régionale de Nancy (MAT_REG_NANCY) à droite. Ils ont chacun une interaction très étroite avec un établissement de troisième ligne avec lequel ils partagent la quasi totalité de leurs patients : le Centre hospitalier d'Épinal (CH_EPINAL) à gauche, le centre Alexis Vautrin (CRLCC_AV_VAN) à droite. La clinique de l'Arc-en-ciel et la Maternité Régionale pratiquent, entre autres, de la chirurgie carcinologique. Ils adressent leurs patients au CH d'Épinal pour l'un, et au Centre Alexis Vautrin pour l'autre, pour de la chimiothérapie et de la radiothérapie car ils ne disposent ni du plateau technique, ni des équipes habilitées à délivrer ces traitements. Leur activité est donc fortement dépendante de la collaboration avec un unique autre centre. Notons, que dans chaque cas, cette interaction a lieu entre hôpitaux situés dans la même ville. Les patients non impliqués dans cette coopération peuvent être vus comme des exceptions, qui ne suivent pas la trajectoire de soins majoritaire pour une raison ou pour une autre. Ils peuvent par exemple être atteints d'une pathologie bien spécifique ou encore choisir un autre établissement de référence pour des convenances personnelles. Au final, la figure illustre bien le fait que les extensions des concepts instables n'ont pas vraiment d'existence réelle en tant que communauté sociale. Ainsi, l'interaction entre la Maternité régionale et le Centre Alexis Vautrin explique bien mieux l'existence de la population de patients traités à la maternité, que l'activité de la maternité seule.

Le second type d'interaction identifié par les concepts fréquents instables est représenté sur la figure 5.11. Il concerne un établissement hautement spécialisé dans le traitement du cancer, de réputation internationale : l'Institut Gustave Roussy (IGR) à Paris. Cet hôpital a reçu 20 patients qui lui ont été adressés par 8 établissements différents de Lorraine. L'IGR joue ici un rôle de recours pour des tumeurs rares et des cas particulièrement complexes. En termes de communauté sociale, les 20 patients sont en fait une multitude de petits groupes relativement distincts, ce qui explique la faible stabilité du concept IGR_PARIS. Ce concept est une illustration du caractère polythétique d'une classe d'objets.

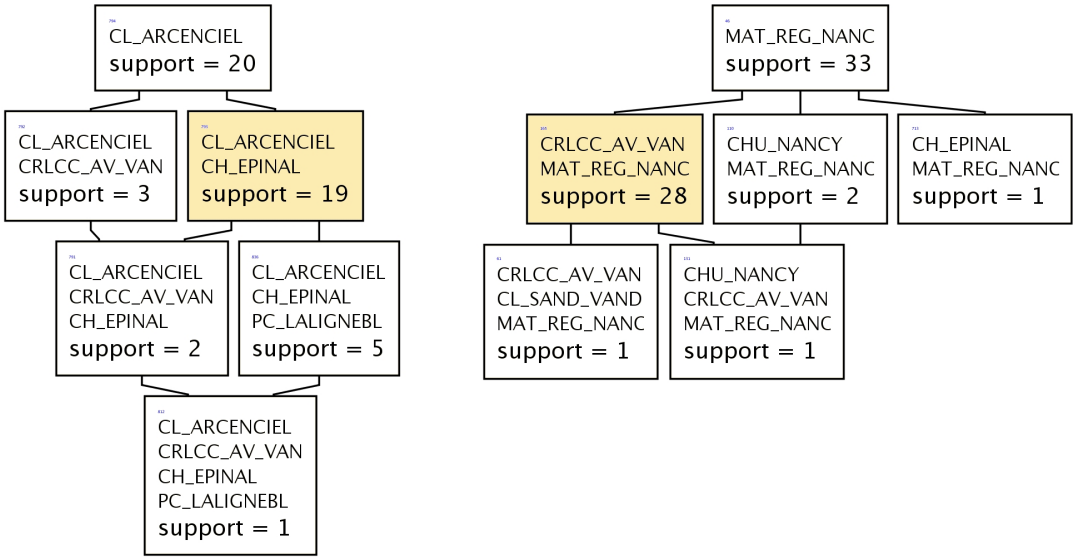

FIG. 5.10 – Concepts fréquents instables : coopérations exclusives.

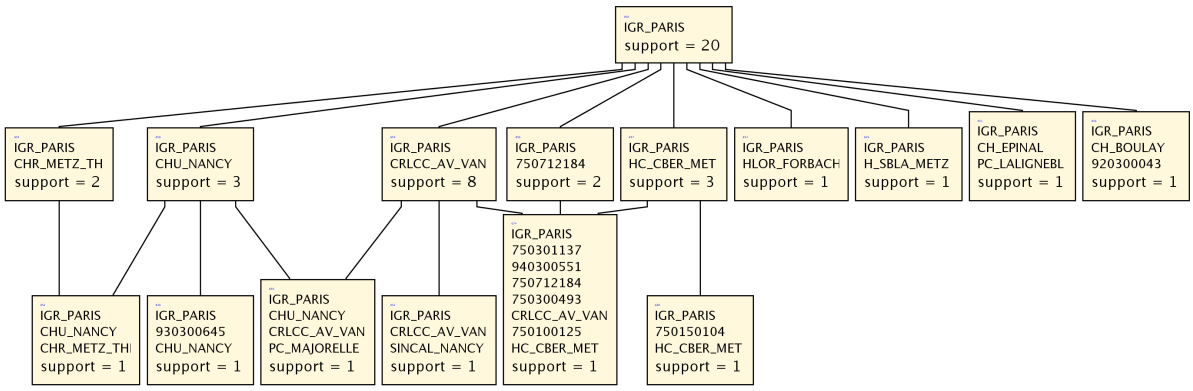

FIG. 5.11 – Concepts fréquents instables et polythétisme : l'IGR a de multiples interactions compartimentées.

Si l'on décidait d'élaguer le treillis en fonction du seuil de stabilité sur la figure 5.10, seul les concepts grisés subsisteraient. Cela peut être souhaitable pour simplifier la visualisation du treillis. Néanmoins, les concepts instables et fréquents peuvent également faire l'objet d'une attention particulière et constituer un objectif d'extraction des connaissances.

5.4.4 Concepts stables et rares

Le quadrant inférieur droit de la figure 5.9 définit sept concepts que l'on peut qualifier de stables et rares. Quatre d'entre eux représentent des interactions entre trois établissements. Ce type d'interaction est logiquement moins fréquent et leur support est dépendant du support des interactions entre deux établissements. Les collaborations entre trois établissements sont les suivantes :

- CHU de Nancy, Centre Alexis Vautrin, CHR de Metz-Thionville.
- CHU de Nancy, Centre Alexis Vautrin, Centre Hospitalier de Toul.
- CHU de Nancy, Centre Alexis Vautrin, Clinique Louis Pasteur à Nancy.
- CHU de Nancy, Polyclinique la Ligne Bleue à Épinal, Centre Hospitalier d' Épinal.

Leur support est de 11 à 12 patients. Elles font toutes intervenir le CHU de Nancy, et trois sur quatre impliquent également le Centre de Lutte Contre le Cancer Alexis Vautrin à Nancy. Ces deux établissements sont ceux qui traitent le plus de cancer en Lorraine. Il est très probable que chacun des établissements soit intervenu pour une phase particulière : chirurgie, radiothérapie et chimiothérapie. Ces interactions sont stables car la probabilité qu'un patient soit hospitalisé dans un quatrième établissement est faible. C'est de plus souhaitable pour la qualité de vie des patients qui ne devraient pas outre mesure être ballottés dans le système de soins.

Les autres concepts stables et rares concernent des interactions entre deux institutions. La figure 5.12 montre la différence entre deux concepts rares, de même support : 11 patients. Les concepts sont étiquetés par leur intension, leur stabilité, et leur support absolu. A gauche, la coopération entre Freyming et Saint Avold est stable et ne se décompose pas ou très peu en faisant intervenir un troisième établissement. A droite, 6 des 11 patients hospitalisés à la fois à Neufchâteau et au CHU de Nancy sont également hospitalisés au Centre de Lutte Contre le Cancer Alexis Vautrin (CAV). De fait, l'interaction entre Nancy et Neufchâteau est moins stable. Une partie de leur coopération s'explique par l'activité d'un troisième établissement. La stabilité permet ainsi de distinguer des concepts qui ne peuvent être discriminés par le support seul. Les concepts rares et stables diffèrent des concepts rares instables en ce sens que leurs attributs suffisent à expliquer la similarité de leurs objets.

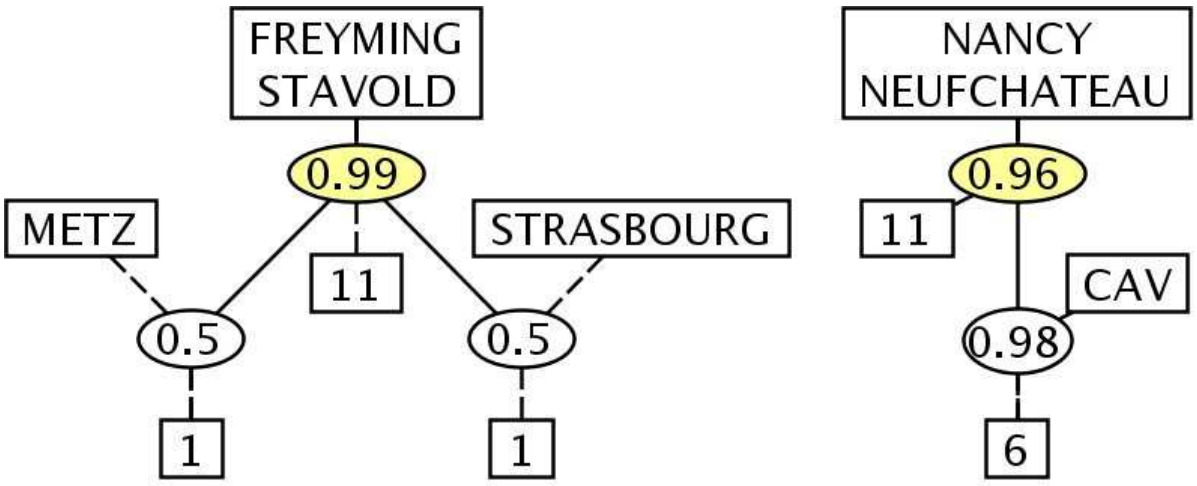

FIG. 5.12 – Deux concepts rares de même support et de stabilités différentes.

5.5 Classification des profils de soins

5.5.1 Problématique

Dans la section 4.3, nous avons introduit la notion de Système de Classification de Malades (SCM). La plupart des SCM regroupent des séjours hospitaliers pour expliquer la variance d'une variable dépendante : coût, survie, mesure de l'état de santé général... Les SCM sont souvent construits par un algorithme de segmentation qui prend en entrée un certain nombre de variables explicatives : durée de séjour, actes médico-chirurgicaux, sexe, âge ... Cet algorithme ne produit pas forcément des classes cliniquement cohérentes et le SCM est raffiné à dire d'experts pour rassembler des problèmes médicaux semblables. L'élaboration des SCM est un processus long et coûteux dont le déploiement nécessite un système d'information dédié. De plus, chaque SCM est bâti pour décrire un type d'activité particulier et manque de souplesse pour être transposé à d'autres domaines, ou décrire les problèmes médicaux à différents niveaux de granularité. Nous nous intéressons ici à la construction d'un système de classification des profils de soins en cancérologie à partir de données existantes, collectées par le PMSI. Le PMSI a été conçu pour classer des hospitalisations dans des hôpitaux de cours séjours. C'est un SCM « généraliste » dont les regroupements n'ont pas une granularité assez fine pour décrire précisément la cancérologie. De plus il est centré sur une hospitalisation et ne permet pas d'appréhender la trajectoire de soins d'un patient dans le temps, à travers une séquence d'évènements médicaux.

Nous avons reconstitué le parcours d'une population de patients opérés pour cancer du sein en 2004 et observés sur une durée de trois ans. Afin de classer ces patients (en fait essentielle-

ment des patientes) par profil pathologique, un contexte formel a été créé avec comme objets les patients, et comme attributs la liste des diagnostics enregistrés lors de leurs hospitalisations entre 2004 et 2006. Le treillis de concepts obtenu permet de distinguer différents profils pathologiques, incluant non seulement la survenue de complications du cancer, mais également la présence de co-morbidités (des pathologies intercurrentes, indépendantes du cancer). L'objectif est de proposer un système de classification ad hoc, non supervisé, capable de produire des classes de patients homogènes sur le plan clinique.

Enfin, nous avons étudié la cohérence entre les concepts obtenus et des variables mesurant le coût des hospitalisations et la mortalité. Le coût a été estimé à partir de l'échelle nationale de coûts [1], un outil permettant d'évaluer la quantité de ressource mobilisées pour chaque hospitalisation. La mortalité a été estimée à partir du mode de sortie d'hospitalisation, une variable du PMSI permettant de repérer les décès survenus à l'hôpital. Ces deux estimations ne sont pas exemptes de biais, mais leur discussion dépasse le propos de ce document.

5.5.2 Treillis des profils de morbidité

Le contexte formel contient 192 patients et 261 diagnostics différents. Son treillis associé compte 788 concepts. Pour simplifier sa visualisation et faciliter la tâche de l'analyste, nous avons élagué le treillis en ne conservant que les vingt concepts les plus stables (en dehors de $\top$ et $\bot$). Le résultat est présenté sur la figure 5.13.

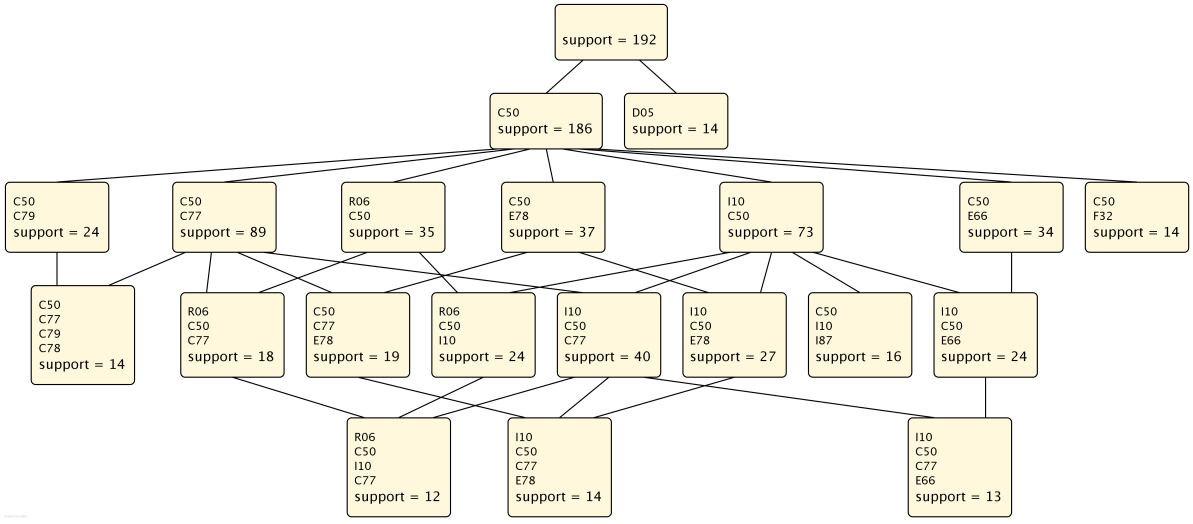

FIG. 5.13 – Treillis des profils de morbidité, élagué par la stabilité.

Les deux premiers concepts après $\top$ distinguent d'emblée les tumeurs malignes infiltrantes

[1] Étude nationale de coûts : http ://www.atih.sante.fr/index.php ?id=0000F00001FF

du sein (C50, 186 patients) des carcinomes in situ du sein (D05, 14 patients). Le carcinome in situ est une forme mineure de cancer du sein, très localisée, qui n'envahit pas le tissu mammaire en profondeur. Son pronostic est bien meilleur et sa prise en charge moins lourde se résume souvent à une exérèse partielle de la glande mammaire. Ceci explique qu'on n'observe pas de sous concept pour le concept D05, car les patients n'ont pas eu d'autres hospitalisations, et donc pas d'autres diagnostics que celui du carcinome in situ initial. De plus la répartition entre cancer in situ et infiltrants est en accord avec les données nationales[Trétarre et al., 2004]. Certains concepts représentent des profils de co-morbidités. 73 patients ont un cancer du sein et une hypertension (C50,I10). Parmi eux, 27 ont également une hypercholestérolémie (C50, I10, E78). D'autres concepts mentionnent une extension secondaire de la tumeur du sein : 89 patients ont un envahissement ganglionnaire (C50, C77), dont 14 ont également des métastases multiples aux systèmes digestif, respiratoire, et autrement spécifié (C50, C77, C78, C79).

La figure 5.13 peut être comparée à la figure 5.14 pour illustrer l'effet de l'élagage du treillis selon la stabilité des concepts. Sur cette dernière figure, nous avons représenté à gauche le filtre du concept (C50, I10, I87). Ce profil regroupe les patients atteints de cancer du sein infiltrant (C50) chez qui on a porté les diagnostics d'hypertension artérielle (I10) et d'insuffisance veineuse périphérique (I87). On remarque que I10 et I87 sont très corrélés à la présence de C50, ce qui vient du mode de sélection de la population (presque tous les patients ont un cancer du sein infiltrant). Par conséquent, les concepts (I10) et (I87), bien que relativement fréquents, sont instables et ne présentent pas d'intérêt dans cette situation. Il en va de même pour le concept (I10, I87). Par ailleurs, les diagnostics I10 et I87 sont également corrélés entre eux : 17 patients sur 20 ayant I87 ont aussi I10. Cette corrélation peut s'expliquer à la fois par des habitudes de codage de l'information médicale de la part de médecins, mais aussi par des caractéristiques épidémiologiques. L'âge moyen des patients ayant une insuffisance veineuse (I87) est de 66 ans, âge auquel on rencontre fréquemment une hypertension artérielle (I10). Ces multiples corrélations entre C50, I10 et I87 entraînent aussi l'instabilité des concepts (C50, I87) et (I10, I87). Ainsi, après élagage par la stabilité, on ne conserve du filtre de (C50, I10, I87) que les quatre concepts à droite de la figure 5.14. Le gain en terme de lisibilité est conséquent, puisque le nombre de concepts est divisé par deux. Pour l'analyste, la perte d'information est minime et le résultat est cohérent avec ses connaissances du domaine. Il n'aurait pas été possible de simplifier de cette manière le treillis avec un élagage par le support, la visualisation du concept (C50, I10, I87) imposant de conserver son filtre en intégralité. L'effet de l'élagage par la stabilité est encore plus marquant ici que dans l'étude de cas précédente (cf section 5.4). Ceci est du à la plus forte corrélation entre les attributs du contexte actuel.

Le tableau 5.2 donne une estimation des coûts d'hospitalisation et de la mortalité observés sur trois ans. Il s'agit de moyennes calculées par patient et par concept. Les résultats concer-

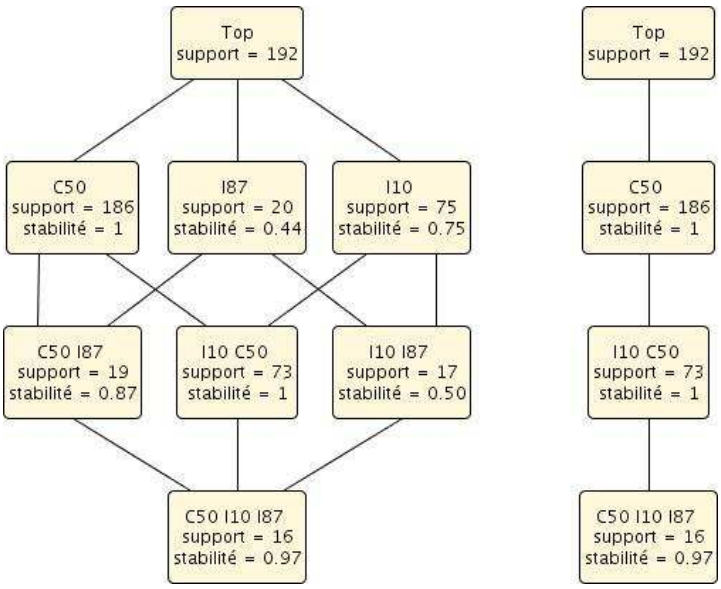

FIG. 5.14 – Effet de l'élagage par la stabilité.

nant l'ensemble de la population étudiée correspondent au concept $\top$ (Top) : le coût moyen total vaut 14 015 €. Les coûts les plus élevés, 29 004 €, sont observés chez les patients au profil pathologique le plus grave : cancer du sein infiltrant avec métastases multiples (concept (C50, C77, C79, C78)). Les coûts les plus faibles correspondent au profil morbide de meilleur pronostic : cancer du sein in situ (D05), 9 793 €. Pour ce dernier, la part de la radiothérapie et de la chimiothérapie est plus faible que dans le profil le plus grave : 25% contre 31%. En effet, le carcinome in situ est souvent traité par chirurgie seule. En général, la présence d'un diagnostic de métastase (C77, C78, C79) est synonyme de surcoût. Il s'agit de stades avancés de la maladie nécessitant des traitements plus longs et plus coûteux. Il en va de même, dans une moindre mesure, pour la présence de co-morbidités comme l'obésité (E66), l'hyperlipidémie (E78) ou l'hypertension (I10). La présence de ces pathologies associées peut compliquer la prise en charge de patients dont l'état général, moins bon, peut contre-indiquer certaines thérapeutiques ou diminuer la résistance à l'agressivité des traitement anticancéreux. Le surcoût des complications et des co-morbidités peut être évalué en comparant un concept avec ses sous-concepts. Ainsi, en comparant les concepts (C50, C77, C79, C78) et (C50, C79), on peut évaluer le surcoût lié à la présence des code diagnostiques C77 (métastase ganglionnaire) et C78 (métastase digestive ou pulmonaire).

Les chiffres observés pour la mortalité sont également cohérents avec les profils de morbidité extraits par le treillis. Le taux le plus fort (43%) est observé pour des stades avancés de la maladie avec métastases multiples (C50, C77, C79, C78), le taux le plus faible pour les

TAB. 5.2 – Coûts et taux de mortalités moyens par concept.

Concept	n	Coûts en €			Taux de mortalité
		Radio et chimiothérapies	Hospitalisations	Total	
(C50, C77, C79, C78)	14	9 092	19 912	29 004	0,43
(C50, C79)	24	6 990	16 776	23 766	0,25
(R06, C50, C77)	18	8 349	12 553	20 902	0,11
(R06, C50, I10, C77)	12	7 637	12 943	20 580	0,08
(C50, C77)	89	7 660	10 644	18 304	0,07
(I10, C50, C77, E66)	13	6 597	11 411	18 008	0,08
(I10, C50, C77, E78)	14	7 696	10 269	17 965	0,00
(I10, C50, C77)	40	6 980	10 704	17 684	0,05
(C50, C77, E78)	19	7 393	10 120	17 513	0,00
(R06, C50, I10)	24	5 094	12 069	17 163	0,08
(R06, C50)	35	5 452	11 512	16 964	0,09
(C50, I10, I87)	16	4 826	11 372	16 198	0,00
(C50, F32)	14	4 778	11 258	16 036	0,07
(I10, C50, E66)	24	4 795	10 531	15 326	0,04
(I10, C50)	73	5 037	10 032	15 069	0,04
(C50, E66)	34	4 876	10 127	15 003	0,03
(I10, C50, E78)	27	5 489	9 079	14 568	0,00
(C50)	186	5 299	9 001	14 300	0,04
Top	192	5 133	8 882	14 015	0,04
(C50, E78)	37	5 192	8 777	13 969	0,00
(D05)	14	2 361	7 432	9 793	0,00

carcinomes in situ (D05). Là encore, la présence de codes de complications ou de co-morbidités est synonyme de surmortalité. Par exemple, le code R06 fait passer la mortalité de 4 à 9% si on compare les concepts (C50) et (R06, C50), et de 9 à 11% si on compare les concepts (R06, C50) et (R06, C50, C77). Le code R06 signifie « dyspnée ». Ce trouble de la respiration peut être le signe d'une atteinte pulmonaire ou cardiaque. Notons enfin noter la surmortalité observée pour le code F32 (épisode dépressif). On peut le voir ici comme un indicateur de la gravité du cancer.

Un treillis de concepts permet d'obtenir une partition de l'ensemble des objets (cf objets propres et treillis d'héritage, section 1.3.3). Or, la structure obtenue après élagage par la stabilité n'est pas forcément un treillis et il n'est pas possible de rattacher un objet à un seul concept du treillis élagué. Un patient peut donc ici être affecté à plusieurs profils de morbidité. Ce n'est pas nécessairement un problème et d'autres systèmes de classification de malades fonctionnent de cette manière comme les Ambulatory Patient Groups [Goldfield et al., 2008]. Néanmoins, les concepts obtenus peuvent être réutilisés comme variable d'entrée pour d'autres méthodes de classification. Nous donnons ici à titre d'exemple un arbre de régression qui explique le coût total des hospitalisations à partir des concepts les plus stables. L'algorithme utilisé est implémenté dans le logiciel R [R Development Core Team, 2007]. L'arbre obtenu est représenté

sur la figure 5.15.

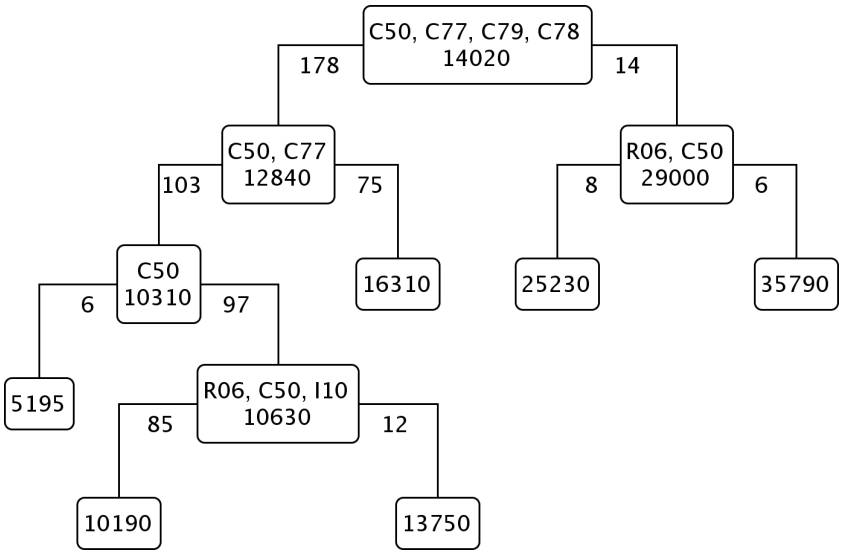

FIG. 5.15 – Arbre de régression du coûts des hospitalisations, expliqués par les concepts stables.

La racine de l'arbre donne le coût moyen pour l'ensemble des patients. L'étiquette des nœuds indique le concept testé et le coût en euros. L'étiquette des branches donne le nombre de patients. Le fils droit correspond à un test positif. Le premier test porte sur le concept (C50, C77, C79, C78), marqueur d'un stade très avancé de cancer. Sa présence est associée à des coûts très élevés. Les codes R06 (dyspnée), C77 (métastase ganglionnaire) et I10 (hypertension artérielle) sont également synonymes de surcoût. Six patients n'ont pas le diagnostic de cancer infiltrant (C50) et présentent les plus faibles coûts d'hospitalisation. La classification obtenue est ainsi cohérente avec les connaissances du domaine.

5.6 Analyse séquentielle des trajectoires de soins

5.6.1 Problématique

Dans les sections précédentes, notre analyse ne tient pas compte du caractère temporel et ordonné des trajectoires de soins. Dans l'histoire d'une maladie, l'ordre des processus pathologiques peut avoir une conséquence sur la prise en charge d'un patient. Inversement, l'ordre des traitements peut modifier l'évolution d'une maladie. Par exemple, la survenue d'une angine de poitrine de novo a un pronostic meilleur qu'après un infarctus du myocarde. Autre exemple, dans le traitement du cancer du sein, la chimiothérapie peut être employée avant (chimiothérapie d'induction) et/ou après une exérèse chirurgicale. Il apparaît donc intéressant d'étudier

l'aspect séquentiel des trajectoires de soins. Pour cela, la recherche de motifs séquentiels fréquents se pose comme une méthode pertinente.

5.6.2 Étude de cas : le cancer du sein

En reprenant la même population que dans la section 5.5, nous nous penchons cette fois sur l'attribut Groupe Homogène de Malades (GHM), enregistré à chaque hospitalisation d'un patient. Rappelons que cet attribut permet de classer les séjours représentant une mobilisation de ressources et un problème médical similaires. Dans la ligne de l'approche adoptée dans la section 5.5, un contexte formel est construit dans lequel les objets sont des patients et les attributs les GHM de leurs différentes hospitalisations. Le treillis dérivé de ce contexte permet d'obtenir une classification des profils de prise en charge des patients. Parmi les concepts les plus stables de ce treillis, nous choisissons pour l'étude de cas présente celui dont l'intension regroupe les GHM suivants :

- 09C05 « Mastectomies subtotales pour tumeur maligne »,
- 24Z02 « Chimiothérapie pour tumeur, en séances »,
- 24Z04 « Autres préparations à une irradiation externe »,
- 24K04 « Mise en place de certains accès vasculaires : séjours de moins de 2 jours ».

Ce concept correspond à 28 patients qui ont eu une chirurgie de la tumeur maligne du sein, de la chimiothérapie et de la radiothérapie. Le GHM 24K04 traduit une hospitalisation pour la pose d'une chambre implantable qui améliore le confort des patients et réduit l'effet nocif des produits anticancéreux sur le site d'injection lorsqu'ils sont administrés par voie veineuse. Nous construisons ensuite la base des séquences contenant la succession des GHM de ces patients. Par exemple, on observe chez l'un d'eux la séquence suivante : $\langle(\texttt{09C05})(\texttt{24K04})(\texttt{24Z02})(\texttt{24Z04})\rangle$. L'objectif de cette étude de cas est de répondre à la question : quelles sont les séquences de soins qui résument le mieux la trajectoire de ces 28 patients ?

Dans un premier temps, une recherche de motifs séquentiels fréquents est menée grâce à l'algorithme slpminer [Seno and Karypis, 2002]. Pour un seuil de 3 patients on retrouve 731 motifs séquentiels fréquents. Le tableau 5.3 montre les 10 plus fréquents d'entre eux. Le motif $\texttt{3} = \langle(\texttt{09C05})(\texttt{24K04})\rangle$ supporté par tous les patients montre que la pose d'une chambre implantable survient toujours après la chirurgie du sein et au cours d'une autre hospitalisation. Le motif $\texttt{7} = \langle(\texttt{09C05})(\texttt{24K04})(\texttt{24Z02})\rangle$ indique que la séquence 3 est suivie de chimiothérapie chez 21 patients. Il montre également que 7 patients qui ne supportent pas ce motif ont eu une pose de chambre implantable non suivie de chimiothérapie, ce qui peut paraître surprenant. Le motif $\texttt{9} = \langle(\texttt{24K04})(\texttt{24Z02})\rangle$ est également supporté par 21 patients. De plus c'est une sous-séquence du motif 7 et il s'agit en fait des mêmes patients. Ces deux motifs illustrent la notion de motif séquentiel fermé présentée dans la section 2.4.3 : le motif 9 n'est pas fermé car

une de ses super-séquences a le même support. On peut faire la même remarque à propos des motifs séquentiels 1 et 5 qui sont supportés par les 28 patients de la base. Ils sont tous deux des sous-séquences du motif 3, également supporté par les 28 patients. La trajectoire commune au 28 patients peut ainsi se résumer aux trois motifs séquentiels fermés 2,3 et 4. La recherche de motifs séquentiels fermés permet donc de réduire le nombre de motifs séquentiels extraits, sans perte d'information.

TAB. 5.3 – Les 10 motifs séquentiels les plus fréquents.

ID	support	motif
1	28	⟨(24K04)⟩
2	28	⟨(24Z04)⟩
3	28	⟨(09C05)(24K04)⟩
4	28	⟨(24Z02)⟩
5	28	⟨(09C05)⟩
6	25	⟨(24Z02)(24Z04)⟩
7	23	⟨(09C05)(24Z02)⟩
8	21	⟨(09C05)(24K04)(24Z02)⟩
9	21	⟨(24K04)(24Z02)⟩
10	20	⟨(24Z07)⟩

Pour simplifier encore un peu plus les résultats, il est possible de construire un nouveau contexte formel avec les patients comme objets et les motifs séquentiels fréquents extraits comme attributs. La relation entre objets et attributs est définie comme suit : un patient p et un motif séquentiel fréquent S sont liés si p supporte S.

Dans le treillis dérivé de ce contexte, les intensions des concepts sont des ensembles de motifs séquentiels fréquents. L'information contenue par ces intensions est généralement redondante : les éléments de l'intension peuvent être réduits aux motifs séquentiels fermés qu'elle contient. Par exemple, un des concepts correspondant à 14 patients a l'intension suivante :

- ⟨(24Z02)⟩
- ⟨(09C05)⟩
- ⟨(24Z06)⟩
- ⟨(24K04)⟩
- ⟨(24Z04)⟩
- ⟨(24Z02)(24Z06)⟩
- ⟨(09C05)(24K04)⟩
- ⟨(09C05)(24Z02)⟩

Parmi ces séquences, les quatre premières ne sont pas fermées et peuvent être déduites des trois dernières.

Le treillis obtenu est encore volumineux et contient 553 concepts. Nous avons représenté un iceberg faisant apparaître les 5% des concepts les plus fréquents dans la figure 5.16. Pour plus de lisibilité, le treillis est orienté de gauche à droite. La notation des séquences est différente : deux parties d'une séquence sont séparées par « -1 ». De plus, dans chaque intension, seuls les motifs séquentiels fermés apparaissent. On peut noter une forte corrélation entre les motifs séquentiels : de nombreux concepts ont un support très proche de celui de leurs sous-concepts. Par exemple, les concepts correspondant aux intensions suivantes (à droite sur la figure) sont très semblables :

- $\{$09C05 −1 24K04 −1 24Z04, 09C05 −1 24Z02 −1 24Z04$\}$
- $\{$09C05 −1 24K04 −1 24Z02, 09C05 −1 24Z02 −1 24Z04$\}$

La différence entre l'extension de ces deux concepts et celle de leur sous-concept commun ($\{$09C05 −1 24K04 −1 24Z02 −1 24Z04$\}$) n'est due qu'à un patient. Dans le cas non séquentiel, les patients de ces trois concepts ne donneraient lieu qu'à un seul concept car ils possèdent tous les attributs de l'ensemble $\{$09C05, 24K04, 24Z02, 24Z04$\}$. Dans le cas séquentiel, de petites variations peuvent donner lieu à de nombreuses séquences semblables et de supports légèrement différents.

La figure 5.17 montre le même treillis de motifs séquentiels élagué cette fois à partir de la stabilité. Les 5% des concepts les plus stables ont été conservés. Contrairement à l'iceberg de la figure 5.16, on voit apparaître des séquences plus longues. Même si leur support est en général plus faible, ces séquences sont intéressantes car elles illustrent des prises en charge plus complètes du cancer du sein. Par exemple les séquences :

- $\{$09C05 −1 24K04 −1 24Z02 −1 24Z04 −1 24Z06$\}$
- $\{$09C05 −1 24K04 −1 24Z02 −1 24Z04 −1 24Z07$\}$

correspondent aux traitements suivants :

- mastectomie (09C05) puis
- mise en place d'un dispositif d'accès vasculaire (24K04) puis
- chimiothérapie (24Z02) puis
- préparation à une radiothérapie (24Z04) puis
- radiothérapie (24Z06 et 24Z07)

Ces séquences sont conformes à la prise en charge standard d'un cancer nécessitant une chirurgie, de la chimiothérapie et de la radiothérapie.

Pour comparer les effets de l'élagage par le support d'une part, et la stabilité d'autre part, nous avons représenté sur la figure 5.18 des extraits des figures 5.16 et 5.17. A gauche, on peut voir l'effet de l'élagage par le support sur les descendants d'un concept et à droite l'effet de l'élagage par la stabilité sur les descendants de ce même concept. Alors que la stabilité fait apparaître le motif séquentiel $\{$09C05 −1 24K04 −1 24Z02 −1 24Z04 −1 24Z07$\}$, le

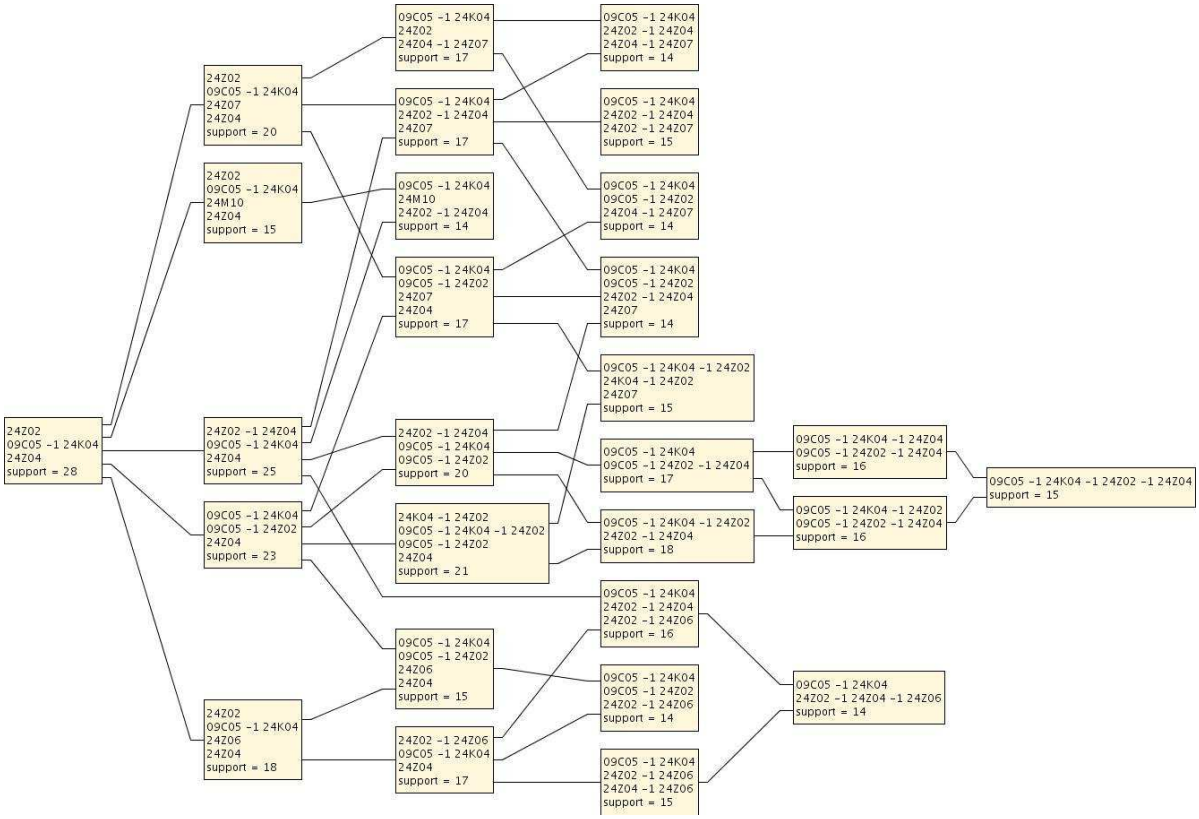

FIG. 5.16 – Iceberg de motifs séquentiels fréquents.

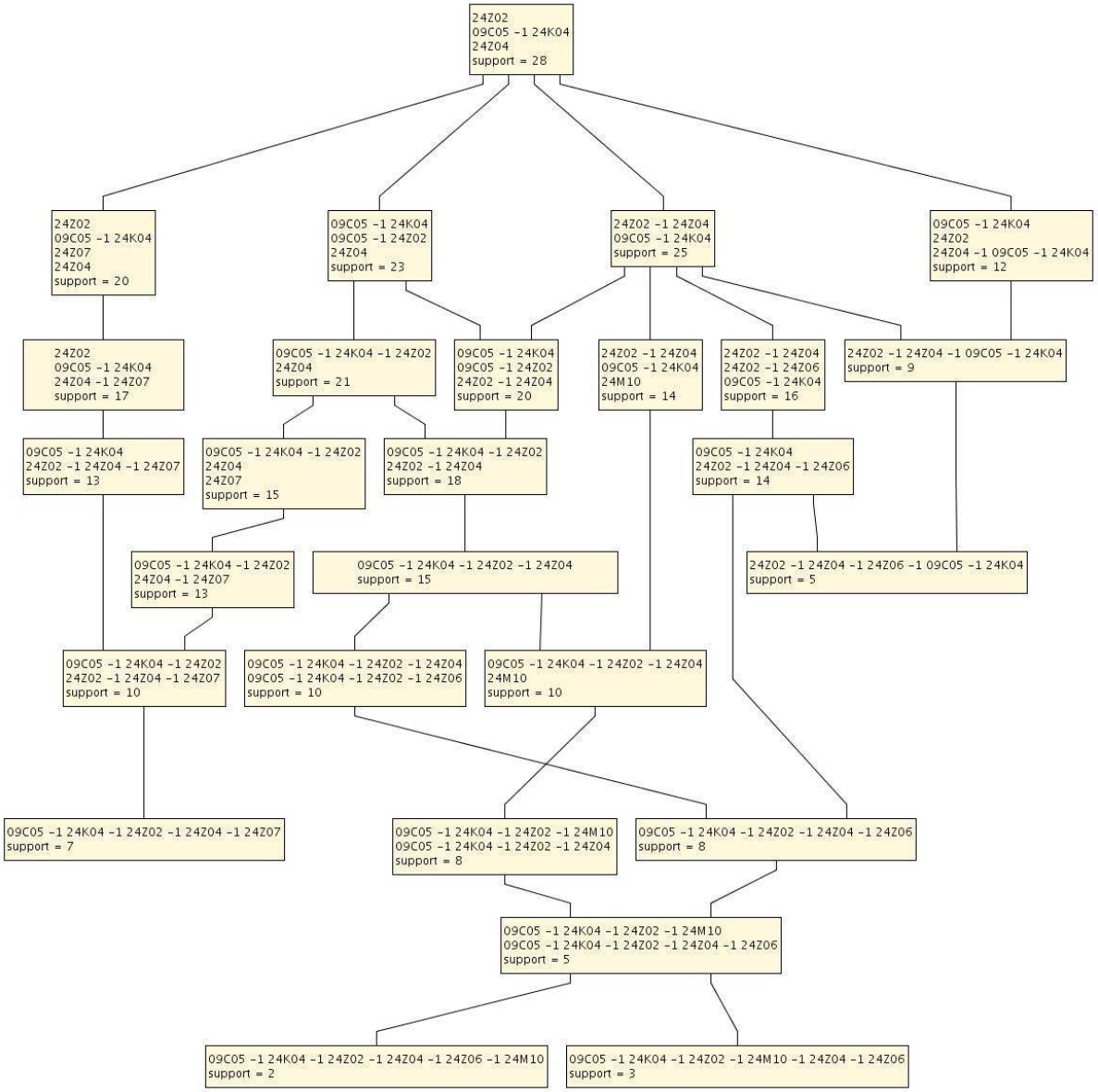

FIG. 5.17 – Treillis de motifs séquentiels élagué par la stabilité.

support ne permet pas de le visualiser. De plus, le support conserve de nombreuses sous-séquences, peu informatives, de ce motif. La stabilité permet de résumer une information quasi identique de manière plus compacte.

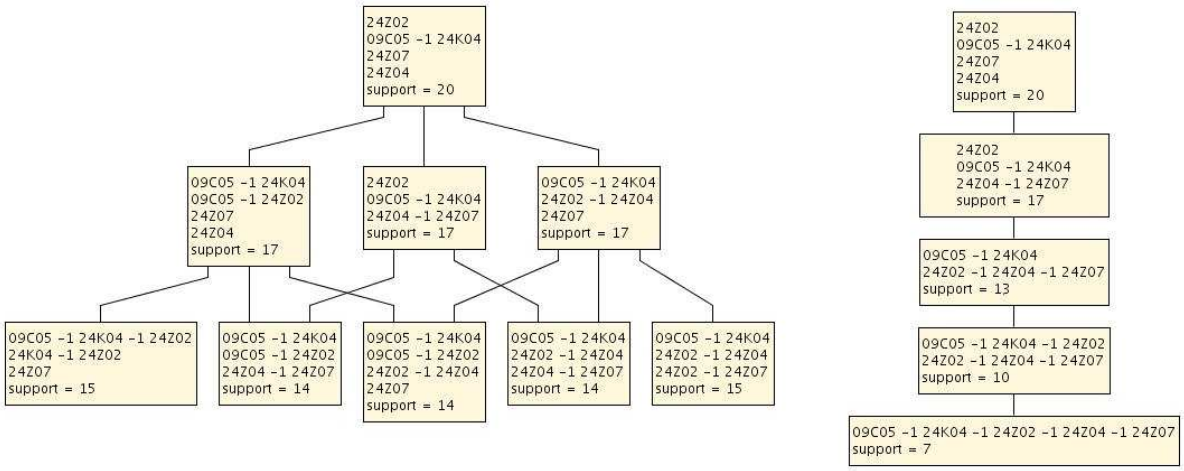

FIG. 5.18 – Comparaison des effets de l'élagage des motifs séquentiels par le support (à gauche) et la stabilité (à droite).

Dans une optique de modélisation des processus de soins, la stabilité semble donc être une mesure d'intérêt plus pertinente que le support pour sélectionner les motifs séquentiels.

5.7 Conclusion

Dans ce chapitre, nous avons montré l'intérêt de l'AFC pour l'étude des trajectoires de soins à partir de données existantes.

Tout d'abord, grâce à ses capacités de visualisation, l'AFC permet de mettre en évidence la topologie de l'organisation virtuelle du système de soins. Contrairement aux graphes utilisés traditionnellement dans l'analyse des réseaux sociaux, l'AFC offre la possibilité d'enrichir la description des liens qui unissent plusieurs acteurs d'un réseau social. Elle permet également de matérialiser des relations entre plus de deux acteurs sous la forme de communautés sociales. La méthode que nous avons élaborée a été utilisée en pratique par les Agences Régionales de l'Hospitalisation de Lorraine et de Champagne-Ardenne. La simplicité de la représentation du fonctionnement du système de soins permet d'éclairer les experts dans leurs prises de décisions stratégiques pour la planification et l'organisation des soins. Cette partie a donné lieu à plusieurs communications [Jay et al., 2005a, Jay et al., 2005b, Jay et al., 2006c, Jay et al., 2006a, Jay et al., 2006b, Jay et al., 2007].

Dans un treillis de grande taille, le support n'est pas toujours la mesure la plus adaptée pour sélectionner les concepts les plus pertinents. Nous avons vu que la stabilité mesure l'intérêt des concepts par une approche différente. Utilisée avec le support, elle permet d'isoler deux types particuliers de concepts : les concepts rares et stables et les concepts fréquents et instables. En terme de fonctionnement du système de soins, ces deux types de concepts illustrent des formes spécifiques d'interactions entre hôpitaux. Cette approche a fait l'objet d'une communication [Jay et al., 2008].

Par un exemple concret, nous avons mis en évidence le potentiel de l'AFC dans l'élaboration d'un système de classification de trajectoires de soins. Bien que non supervisée, la classification par AFC détermine des profils de morbidité qui présentent une cohérence clinique. Elle constitue en cela une alternative intéressante aux systèmes actuels de classification de patients. En effet, ceux-ci manquent souvent de souplesse pour être adaptés à de nouveaux objectifs. L'AFC laisser envisager la possibilité de créer à moindre coût des systèmes de classification ad hoc, à partir de données existantes, ne nécessitant pas la mise en place d'un nouveau recueil d'information.

Enfin, nous avons abordé la question de l'exploration de données séquentielles par AFC. Dans ce domaine, le volume des résultats que l'analyste doit traiter à l'issue du processus d'ECD représente clairement une limite due à des problèmes combinatoires. Cette limite est partiellement repoussée par l'extraction de motifs séquentiels fermés. Nous avons illustré par une étude de cas la capacité de la stabilité à simplifier encore les résultats de la recherche de motifs séquentiels fréquents. Il s'agit là d'une piste de recherche qui nécessite d'être approfondie notamment pour des questions de mise à l'échelle. En effet, le calcul de la stabilité reste un problème #P-complet qui ne peut être simplifié qu'après le calcul du treillis de concepts dans son intégralité.

Conclusion

Contributions

La connaissance et la maîtrise des trajectoires de soins par les professionnels de santé est un facteur d'amélioration de la qualité et de l'efficience des soins. Dans une approche fondée sur l'extraction de connaissances à partir de données, nous avons présenté un cadre général d'analyse des trajectoires de soins par treillis de concepts.

Dans un premier chapitre, nous avons introduit les bases théoriques de l'Analyse Formelle de Concepts et rappelé ses principaux domaines d'utilisation. Dans le chapitre suivant, nous avons décrit l'AFC en tant que méthode d'extraction des connaissances à partir de données en retraçant ses liens étroits avec la recherche de motifs fréquents. Plusieurs raisons expliquent l'intérêt de l'AFC en extraction des connaissances. Elle constitue un cadre théorique qui facilite la manipulation des motifs fréquents :

- lors de leur extraction. L'AFC est à la base d'une famille d'algorithmes efficaces de recherche de motifs fréquents.
- Pour leur interprétation. L'AFC permet de réduire le nombre de motifs extraits sans perte d'information. De plus elle offre la possibilité de les classer et de les visualiser.

Une des principales limites de l'AFC est liée à la taille des données et pose plusieurs problèmes :

- le calcul d'un treillis de concepts est #P-complet.
- La visualisation perd de son intérêt pour les treillis de grande taille.
- il est indispensable de trouver un compromis entre utilité des connaissances véhiculées par les concepts et nombre de concepts à traiter par l'analyste.

Dans le troisième chapitre, nous nous sommes penchés sur cette dernière question. Les mesures de qualité des concepts pour réduire la complexité des treillis de grande taille sont un sujet encore relativement peu abordé. Même si l'AFC est de plus en plus utilisée en extraction des connaissances, les concepts formels ne sont pas toujours l'objectif final du processus, mais

plutôt une étape intermédiaire comme c'est le cas par exemple dans la génération de règles d'association. Pourtant, dans certains domaines tels que l'analyse de communautés sociales, les concepts formels constituent véritablement les unités de connaissances que l'on cherche à mettre en évidence. A ce jour, il n'existe à notre connaissance que deux mesures d'intérêt des concepts formels : le support et la stabilité. Dans une analyse qualitative détaillée, nous nous sommes attachés à démontrer la capacité de la stabilité à isoler des concepts intéressants. La stabilité peut d'abord être vue comme une mesure de cohésion des concepts et permet de distinguer les concepts de caractère monothétique des concepts de caractère polythétique. Cette propriété s'avère très intéressante pour l'analyse des communautés sociales. La stabilité donne un point de vue différent là où le support échoue à révéler de petites communautés à forte cohésion. Ce point de vue tranche avec l'hypothèse souvent émise que des connaissances potentiellement intéressantes touchent à des phénomènes fréquents. Toutefois, la stabilité peut être utilisée conjointement avec le support pour extraire des concepts particuliers :

- les concepts rares et stables.
- les concepts fréquents et instables.

Après une description du contexte médical, nous avons mis en œuvre dans le cinquième chapitre une approche fondée sur l'AFC pour analyser les trajectoires de soins. À partir de données réelles issues d'un système d'information hospitalier français, le PMSI, nous avons proposés diverses méthodes pour établir des classifications de trajectoires selon plusieurs points de vue. Le premier exemple montre que les treillis de concepts sont particulièrement bien adaptés à la visualisation de l'organisation virtuelle d'un réseau de soins. Par rapport aux graphes traditionnellement utilisés en analyse de réseaux sociaux, les treillis de concepts permettent de révéler et hiérarchiser des liens complexes entre acteurs du réseau. Dans notre domaine d'analyse, ils nous ont permis par exemple de mettre en évidence des interactions entre trois hôpitaux pour traiter des patients communs. De plus la stabilité et le support peuvent être utilisés pour caractériser ces liens en mesurant la cohésion et l'intensité des coopérations entre hôpitaux. Ces mesures constituent également un moyen de réduire la complexité des treillis volumineux. Enfin, la structure hiérarchique des treillis peut servir de point de départ à l'élaboration d'un système d'interrogation et de navigation des trajectoires de soins. Toutes ces propriétés font de l'AFC un véritable outil d'évaluation et d'aide à la décision pour les responsables et les acteurs du système de santé.

Dans un deuxième exemple, nous avons voulu établir une classification des profils pathologiques. Il s'agit là d'un objectif classique en traitement de l'information médicale lorsque l'on cherche à réduire l'infinie diversité des problèmes de santé en un petit nombre de classes. Nous avons montré que, bien que non supervisée, l'AFC est une méthode de classification capable de regrouper des malades en classes cliniquement cohérentes. Notre approche se distingue

de la plupart des systèmes de classification de malades fondés sur une analyse de la variance d'un variable cible, et raffinée par des experts. Ces systèmes souffrent d'une certaine rigidité et s'adaptent difficilement à un changement de granularité ou d'objectif. Nous pensons que l'AFC est un moyen relativement peu coûteux permettant de réutiliser les données créées par ces systèmes d'information afin d'obtenir des classifications ad hoc. Dans cet exemple, nous avons par ailleurs à nouveau mis en évidence l'intérêt de la stabilité pour réduire la complexité de treillis volumineux et isoler les concepts les plus pertinents.

Enfin, nous nous sommes penchés sur l'aspect séquentiel des trajectoires de soins. Bien que la recherche de motifs fréquents séquentiels semble une méthode adaptée à ce type d'analyse, elle pose des difficultés d'interprétation des résultats dus à des phénomènes combinatoires encore plus notables que dans le cas non séquentiel. Les solutions envisagées jusqu'alors pour réduire la taille des résultats reposent sur le support et la notion de motifs séquentiels fermés. Ici, la définition de fermeture est d'ailleurs sensiblement différentes de celle qui est utilisée en AFC. Dans un dernier exemple, nous avons essayé de montrer qu'une classification des motifs séquentiels par treillis de concepts, simplifiée par la stabilité, permet de d'extraire des types de trajectoires qui représentent des séquences plus complètes que ne l'aurait autorisé le support seul. Au total, il ressort de notre expérience que la stabilité peut être utilisée pour élaguer un treillis de concepts en largeur, alors que le support correspond à un élagage en profondeur. Dans le cas séquentiel où il existe une forte similarité entre l'extension d'un concept et celles de ses sous-concepts, la stabilité semble donc plus intéressante.

En conclusion, l'AFC s'est avéré être une méthode de classification singulièrement intéressante pour l'analyse des trajectoires de soins. Elle se distingue surtout par sa souplesse et la diversité des problèmes qu'elle est capable de traiter. À l'origine limitée par la taille des données, son utilisation dans le cadre de l'extraction des connaissances bénéficie des apports des mesures de qualité des concepts que sont le support et la stabilité. Si notre domaine d'étude se limite aux trajectoires de soins, notre approche est aisément transposable à d'autres champs d'application de l'extraction de connaissances à partir de données symboliques et notamment l'analyse de réseaux sociaux, qu'ils soient scientifiques, culturels, économiques, d'animaux, d'internautes...

Perspectives

Les trajectoires de soins sont des objets relativement complexes. Dans ce document, elles sont modélisées assez simplement : ensembles d'hôpitaux fréquentés par un patient, ensembles de diagnostiques ou d'actes médicaux. La construction de séquences d'épisodes de soins constitue cependant déjà un début de structuration des trajectoires. Or, les bases de données comme le PMSI comportent de nombreux attributs que nous n'avons pas exploités. De plus, ces bases

ont une structure relationnelle. Il existe donc une dépendance entre attributs qu'un contexte formel ne permet pas de modéliser de façon satisfaisante. Par exemple, un patient peut être hospitalisé dans un établissement `h1` pour un problème `p` et recevoir les traitements `t1` et `t2`. Il peut être ultérieurement hospitalisé dans un établissement `h2` pour le même problème `p` et y recevoir les traitements `t3` et `t4`. Notre approche ne permet pas à ce stade d'extraire les trajectoires du type : hospitalisation dans `h1` pour un problème `p` et dans `h2` pour le même problème mais avec des traitements différents. Se pose alors la question de la découverte de connaissances à partir de données structurées ou semi-structurées. Ce problème a déjà fait l'objet de nombreux travaux [Zaki et al., 2005], mais on peut s'interroger sur le potentiel de l'AFC pour traiter ce type de données. L'approche adoptée par l'analyse relationnelle de concepts semble prometteuse [Hacene et al., 2007b] mais souffre de limites de mise à l'échelle. Il serait intéressant dans ce cadre d'étudier les propriétés des mesures de qualités des concepts.

Néanmoins, il faut s'attendre à des difficultés liées à la complexité du calcul des ces mesures, notamment de la stabilité qui, à l'inverse du support, ne possède pas la propriété d'antimonotonie permettant d'optimiser les algorithmes de construction d'icebergs de concepts. Il y a donc ici une autre voie de recherche à explorer.

En ce qui concerne la sélection de motifs pertinents, nous avons vu que la stabilité était capable de discriminer des concepts rares, indifférentiables par leur support. Le problème de l'extraction de motifs rares a été introduit par Szathmary [Szathmary et al., 2007]. Les domaines d'applications potentiels de cette approche sont nombreux, en particulier en médecine : recherche d'effets indésirables des médicaments, étude de facteurs de risques de maladies rares, analyse des interactions entre génotype et facteurs environnementaux sur l'état de santé... D'autre part, des travaux récents ont présenté l'AFC comme une généralisation de l'analyse des tableaux de contingence [Pogel and Ozonoff, 2008], très utilisés en épidémiologie. Les auteurs proposent d'annoter les concepts du treillis par des mesures d'association comme les Odds Ratios. Plusieurs développement sont ici envisageables. La sélection des concepts pertinents peut sans doute s'appuyer sur les nombreuses mesures d'association développées dans tant le domaine des règles d'association qu'en statistiques traditionnelles. Inversement, on peut se demander si l'AFC peut servir de cadre théorique aux modèles statistiques comme la régression logistique par exemple. De manière plus générale, il serait intéressant de rapprocher l'AFC avec des méthodes plus numériques.

Enfin, le processus d'extraction des connaissances à partir de données comporte une phase d'intégration et de réutilisation des unités extraites. Il est donc indispensable de les maintenir dans un système de représentation des connaissances. Le lien avec les ontologies peut être fait ici à deux niveaux. D'une part, l'AFC est susceptible de supporter la création d'ontologies des trajectoires de soins guidée par les données. En ce sens, les exemples que nous avons présentés

sur les profils de morbidité et les séquences de soins dans les sections 5.5 et 5.6 pourraient servir de point de départ à ce processus. D'autre part, le processus d'analyse des trajectoires de soins pourrait être assisté par la mobilisation de connaissances existantes. Par exemple, le PMSI collecte des informations codées dans des nomenclatures standardisées comme la Classification Commune des actes Médicaux (CCAM). Or la CCAM est issue du modèle de représentation des connaissances GALEN [Trombert-Paviot et al., 2000]. La richesse de ce formalisme est une opportunité à saisir pour optimiser le processus de fouille de données et permettre à la fois un enrichissement sémantique et un meilleur ciblage des connaissances à extraire.

Bibliographie

[Abidi, 2001] Abidi, S. S. (2001). Knowledge management in healthcare : towards 'knowledge-driven' decision-support services. *Int J Med Inform*, 63(1-2) :5–18. 76

[Agrawal et al., 1993] Agrawal, R., Imielski, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *proceedings of the ACM SIGMOD Int'l Conference on Management of Data*, pages 207–216. 39, 40, 47

[Agrawal and Srikant, 1994] Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. In *proceedings of the 20th Int'l Conference on Very Large Data Bases*, pages 478–499. Expanded version in IBM Research Report RJ9839. 42

[Agrawal and Srikant, 1995] Agrawal, R. and Srikant, R. (1995). Mining sequential patterns. In Yu, P. S. and Chen, A. S. P., editors, *Eleventh International Conference on Data Engineering*, pages 3–14, Taipei, Taiwan. IEEE Computer Society Press. 42

[Argyris and Schön, 2002] Argyris, C. and Schön, D. (2002). *Apprentissage organisationnel. Théorie, méthode, pratique.* DeBoeck Université. 63

[Baader and Nutt, 2003] Baader, F. and Nutt, W. (2003). Basic description logics. In Baader, F., Calvanese, D., McGuinness, D. L., Nardi, D., and Patel-Schneider, P. F., editors, *Description Logic Handbook*, pages 43–95. Cambridge University Press. 33

[bailey, 1973] bailey, K. (1973). Monothetic and polythetic typologies and their relation to conceptualization, measurement and scaling. *American Sociological Review*, 38(1) :18–33. 57

[Barbut and Monjardet, 1970] Barbut, M. and Monjardet, B. (1970). *Ordre et Classification*, volume II. Hachette. 7, 13

[Basili et al., 1997] Basili, R., Pazienza, M. T., and Vindigni, M. (1997). Corpus-driven unsupervised learning of verb subcategorization frames. In Lenzerini, M., editor, *AI*IA*, volume 1321 of *Lecture Notes in Computer Science*, pages 159–170. Springer. 36

[Bastide et al., 2000] Bastide, Y., Taouil, R., Pasquier, N., Stumme, G., and Lakhal, L. (2000). Mining frequent patterns with counting inference. *SIGKDD Explor. Newsl.*, 2(2) :66–75. 45, 46

[Bayardo, 1998] Bayardo, R. (1998). Efficiently mining long patterns from databases. In *Proceeding of the 1998 ACM-SIGMOD international conference on management of data (SIGMOD'98)*, page 85–93. 45

[Bendaoud et al., 2007a] Bendaoud, R., Hacene, M. R., Toussaint, Y., Delecroix, B., and Napoli, A. (2007a). Text-based ontology construction using relational concept analysis. In Flouris, G. and d'Aquin, M., editors, *Proceedings of the International Workshop on Ontology Dynamics*, pages 55–68, Innsbruck (Austria). 23

[Bendaoud et al., 2007b] Bendaoud, R., Hacène, M. R., Toussaint, Y., Delecroix, B., and Napoli, A. (2007b). Construction d'ontologie à partir de corpus de textes avec l'acf. In Trichet, F. and Fiorino, H., editors, *Ingénierie des connaissances (IC-2007), Grenoble*, pages 121–132. Cépadues éditions, Toulouse. 33

[Bergson, 1907] Bergson, H. (1907). *L'évolution créatrice*. Les Presses universitaires de France. 26

[Birkhoff, 1940] Birkhoff, G. (1940). *Lattice Theory, 3rd ed.* American Mathematical Society, Providence, Rhode Island. 7

[Bleser et al., 2006] Bleser, L. D., Depreitere, R., Waele, K. D., Vanhaecht, K., Vlayen, J., and Sermeus, W. (2006). Defining pathways. *J Nurs Manag*, 14(7) :553–563. 66, 67

[Boisot, 1995] Boisot, M. (1995). *Information space : a framework for learning in organizations, intitutions and culture*. CENGAGE Lrng Business Press. 63

[Bordat, 1986] Bordat, J. (1986). Calcul pratique du treillis de galois d'une correspondance. *Math. Sci. Hum.*, 96 :31–47. 23

[Brewster et al., 1985] Brewster, A. C., Karlin, B. G., Hyde, L. A., Jacobs, C. M., Bradbury, R. C., and Chae, Y. M. (1985). Medisgrps : a clinically based approach to classifying hospital patients at admission. *Inquiry*, 22(4) :377–387. 69

[Carpineto and Romano, 1998] Carpineto, C. and Romano, G. (1998). Effective reformulation of boolean queries with concept lattices. In *FQAS '98 : Proceedings of the Third International Conference on Flexible Query Answering Systems*, pages 83–94, London, UK. Springer-Verlag. 33

[Carpineto and Romano, 2005] Carpineto, C. and Romano, G. (2005). Using concept lattices for text retrieval and mining. In [Ganter et al., 2005], pages 161–179. 33

[Caspard et al., 2007] Caspard, N., Leclerc, B., and Monjardet, B. (2007). *Ensembles ordonnés finis : concepts, résultats et usages.* Springer Verlag. 14

[Chein, 1969] Chein, M. (1969). Algorithme de recherche des sous-matrices premières d'une matrice. *Bull. Math. Soc. Sci. Math. R.S. Roumanie*, 13 :21–25. 23

[Cherfi, 2004] Cherfi, H. (2004). *Étude et réalisation d'un système d'extraction de connaissances à partir de texte.* PhD thesis, Université Henri Poincaré – Nancy 1. 51

[Cimiano et al., 2005] Cimiano, P., Hotho, A., and Staab, S. (2005). Learning concept hierarchies from text corpora using formal concept analysis. *J. Artif. Intell. Res. (JAIR)*, 24 :305–339. 32

[Cimiano et al., 2004] Cimiano, P., Hotho, A., Stumme, G., and Tane, J. (2004). Conceptual knowledge processing with formal concept analysis and ontologies. In [Eklund, 2004], pages 189–207. 31

[Cole and Stumme, 2000] Cole, R. and Stumme, G. (2000). Cem - a conceptual email manager. In *ICCS '00 : Proceedings of the Linguistic on Conceptual Structures*, pages 438–452, London, UK. Springer-Verlag. 34

[Cristofor et al., 2000] Cristofor, D., Cristofor, L., and Simovici, D. (2000). Galois connections and data mining. *Journal of Universal Computer Science*, 6(1) :60–73. 46

[D'Aquin et al., 2004] D'Aquin, M., Brachais, S., Lieber, J., and Amedeo, N. (2004). Decision support and knowledge management in oncology using hierarchical classification. *Stud Health Technol Inform*, 101 :16–30. 62

[Dart et al., 2003] Dart, T., Cui, Y., Chatellier, G., and Degoulet, P. (2003). Analysis of patient flows using data-mining. In et al., R., editor, *The New Navigators : from Professionals to Patients. Proceedings of Medical Informatics Europe 2003*, pages 263–268. 71

[Davis et al., 1993] Davis, R., Shrobe, H., and Szolovits, P. (1993). What is a knowledge representation ? *AI Magazine*, 14(1) :17–33. 30

[de Luc et al., 2001] de Luc, K., Kitchiner, D.and Layton, A., Morris, E., Murray, Y., and Overill, S. (2001). *Developing Care Pathways : the Handbook.* Radcliffe Medical Press, Oxon. 65

[DeZell et al., 1987] DeZell, A. V., Comeau, E., and Zander, K. (1987). Nursing case management : managed care via the nursing case management model. *NLN Publ*, 20(2191) :253–264. 65

[Duff et al., 2001] Duff, F. L., Happe, A., Burgun, A., Levionnois, S., Bremond, M., and Beux, P. L. (2001). Sharing medical data for patient path analysis with data mining method. *Stud Health Technol Inform*, 84(Pt 2) :1364–1368. 71

[Dufour et al., 2006] Dufour, J.-C., Bouvenot, J., Ambrosi, P., Fieschi, D., and Fieschi, M. (2006). Textual guidelines versus computable guidelines : a comparative study in the framework of the presguid project in order to appreciate the impact of guideline format on physician compliance. *AMIA Annu Symp Proc*, pages 219–223. 62

[Duquenne, 1996a] Duquenne, V. (1996a). Lattice analysis and the representation of handicap associations. *Social Networks*, 18 :217–230. 78

[Duquenne, 1996b] Duquenne, V. (1996b). On lattice approximations : Syntactic aspects. *Social Networks*, 18 :189–199. 78

[Düwel and Hesse., 1998] Düwel, S. and Hesse., W. (1998). Identifying candidate objects during system analysis. In Proceedings of CAiSE'98/IFIP 8.1 Third International Workshop on Evaluation of Modelling Methods in System Analysis and Design (EMMSAD'98), Pisa. 36

[Eklund, 2004] Eklund, P. W., editor (2004). *Concept Lattices, Second International Conference on Formal Concept Analysis, ICFCA 2004, Sydney, Australia, February 23-26, 2004, Proceedings*, volume 2961 of *Lecture Notes in Computer Science*. Springer. 115, 122

[Eklund and II, 2002] Eklund, P. W. and II, R. J. C. (2002). Structured ontology and information retrieval for email search and discovery. In Hacid, M.-S., Ras, Z. W., Zighed, D. A., and Kodratoff, Y., editors, *ISMIS*, volume 2366 of *Lecture Notes in Computer Science*, pages 75–84. Springer. 26

[Elghazel, 2007] Elghazel, H. (2007). *Classification et Prévision des Données Hétérogènes : Application aux Trajectoires et Séjours Hospitaliers*. PhD thesis, Université Claude Bernard Lyon 1. 71

[Fayyad et al., 1996] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). The kdd process for extracting useful knowledge from volumes of data. *Communication of the ACM*, 29(11) :27–34. 29, 30

[Feder et al., 1999] Feder, G., Eccles, M., Grol, R., Griffiths, C., and Grimshaw, J. (1999). Clinical guidelines : using clinical guidelines. *BMJ*, 318(7185) :728–730. 62

[Ferré, 2002] Ferré, S. (2002). *Systèmes d'information logiques : un paradigme logico-contextuel pour interroger, naviguer et apprendre*. PhD thesis, Université de Rennes 1. 26

[Fetter et al., 1980] Fetter, R., Shin, Y., Freeman, J., Averill, R., and JDThompson (1980). Case mix definition by diagnosis-related groups. *Med Care*, 18(2) :1–53. 69

[Fox et al., 1998] Fox, J., Johns, N., and Rahmanzadeh, A. (1998). Disseminating medical knowledge : the proforma approach. *Artif Intell Med*, 14(1-2) :157–181. 62

[Freeman and White, 1993] Freeman, L. C. and White, D. R. (1993). Using galois lattices to represent network data. *Sociological methodology*, 23 :127–146. 77, 78

[Ganter, 1984] Ganter, B. (1984). Two basic algorithms in concept analysis. FB4-Preprint 831, TH Darmstadt. 23

[Ganter and Kuznetsov, 2000] Ganter, B. and Kuznetsov, S. O. (2000). Formalizing hypotheses with concepts. In *ICCS '00 : Proceedings of the Linguistic on Conceptual Structures*, pages 342–356, London, UK. Springer-Verlag. 31

[Ganter et al., 2005] Ganter, B., Stumme, G., and Wille, R., editors (2005). *Formal Concept Analysis, Foundations and Applications*, volume 3626 of *Lecture Notes in Computer Science*. Springer. 114, 120, 122, 123

[Ganter and Wille, 1999] Ganter, B. and Wille, R. (1999). *Formal Concept Analysis, Mathematical foundations*. Springer, Berlin. 8, 18, 19, 23

[Godin et al., 1989] Godin, R., Gecsei, J., and Pichet, C. (1989). Design of a browsing interface for information retrieval. In Belkin, N. J. and van Rijsbergen, C. J., editors, *SIGIR*, pages 32–39. ACM. 13, 33

[Godin and Mili, 1993] Godin, R. and Mili, H. (1993). Building and maintaining analysis-level class hierarchies using galois lattices. In *OOPSLA*, pages 394–410. 36

[Godin et al., 1995a] Godin, R., Mineau, G., Missaoui, R., and Mili, H. (1995a). Méthodes de classification conceptuelle basées sur les treillis de galois et applications. *Revue d'Intelligence Artificielle*, 9 :105–137. 16

[Godin et al., 1995b] Godin, R., Missaoui, R., and Alaoui, H. (1995b). Incremental concept formation algorithms based on galois (concept) lattices. *Computational Intelligence*, 11 :246–267. 23

[Godin et al., 1986] Godin, R., Saunders, E., and Gecsei, J. (1986). Lattice model of browsable data spaces. In *Information Sciences*, volume 40, pages 89–116, New York, NY, USA. Elsevier Science Inc. 13, 33

[Goldfield et al., 2008] Goldfield, N., Averill, R., Eisenhandler, J., and Grant, T. (2008). Ambulatory patient groups, version 3.0–a classification system for payment of ambulatory visits. *J Ambul Care Manage*, 31(1) :2–16. 98

[Gonnella et al., 1984] Gonnella, J. S., Hornbrook, M. C., and Louis, D. Z. (1984). Staging of disease. a case-mix measurement. *JAMA*, 251(5) :637–644. 69

[Green, 1978] Green, P. (1978). An aid/logit procedure for analyzing large multiway contingency tables. *J. Marketing Res.*, 15(1) :132–136. 69

[Gruber, 1995] Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum.-Comput. Stud.*, 43(5-6) :907–928. 31

[Grémy, 1997] Grémy, F. (1997). Vers l'organisation et la coordination du système de soins. *Gestions Hospitalières*, Juin-Juillet :433–438. 68

[Guillet and Hamilton, 2007] Guillet, F. and Hamilton, H. J. (2007). *Quality Measures in Data Mining (Studies in Computational Intelligence)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA. 51

[Guénoche, 1990] Guénoche, A. (1990). Construction du treillis de galois d'une relation binaire. *Math. Inf. Sci. Hum.*, 109 :41–53. 23

[Habermas, 1981] Habermas, J. (1981). *Theorie des kommunikativen Handelns*. Suhrkamp. 29, 30

[Hacene et al., 2007a] Hacene, M. R., Dao, M., Huchard, M., and Valtchev, P. (2007a). Analyse formelle de données relationnelles pour la réingénierie des modèles uml. In Borne, I., Crégut, X., Ebersold, S., and Migeon, F., editors, *LMO*, pages 151–166. Hermès Lavoisier. 23, 36

[Hacene et al., 2007b] Hacene, M. R., Huchard, M., Napoli, A., and Valtchev, P. (2007b). A proposal for combining formal concept analysis and description logics for mining relational data. In Kuznetsov, S. O. and Schmidt, S., editors, *ICFCA*, volume 4390 of *Lecture Notes in Computer Science*, pages 51–65. Springer. 22, 23, 31, 33, 110

[Hamelin, 1925] Hamelin, O. (1925). *Essai sur les éléments principaux de la représentation*. F.Alcan. 26

[Han et al., 2007] Han, J., Cheng, H., Xin, D., and Yan, X. (2007). Frequent pattern mining : current status and future directions. *Data Min. Knowl. Discov.*, 15(1) :55–86. 42

[Hornbrook, 1982] Hornbrook, M. C. (1982). Hospital case mix : its definition, measurement and use : Part i. the conceptual framework. *Med Care Rev*, 39(1) :1–43. 68

[Hornbrook, 1993] Hornbrook, M. C. (1993). Definition and measurement of episode of care in clinical and economic studies. Technical report, Center for health research, Kaiser Permanente, Portland. 68

[Hornbrook and A., 1985] Hornbrook, M. C. and A., M. (1985). Health care episode : definition, measurement and use. *Med Care Rev*, 42 :163–217. 68

[Jay et al., 2005a] Jay, N., Fresson, J., De Gasperi, M., Lacour, B., Mayeux, D., and Kohler, F. (2005a). Analyse de la trajectoire de soins en cancérologie. extraction de connaissances à partir du pmsi. *Journal d'Economie Médicale*, 23(3-4) :191–194. 105

[Jay et al., 2005b] Jay, N., Fresson, J., De Gasperi, M., Lacour, B., Mayeux, D., and Kohler, F. (2005b). Analyse de la trajectoire de soins en cancérologie. extraction de connaissances à partir du pmsi. In *Actes des XVIIIe Journées Emois*. 105

[Jay et al., 2006a] Jay, N., Kohler, F., and Napoli, A. (2006a). Using formal concept analysis for mining and interpreting patient flows within a healthcare network. In Yahia, S. B., Nguifo, E. M., and Belohlávek, R., editors, *CLA*, volume 4923 of *Lecture Notes in Computer Science*, pages 263–268. Springer. 105

[Jay et al., 2008] Jay, N., Kohler, F., and Napoli, A. (2008). Analysis of social communities with iceberg and stability-based concept lattices. In [Medina and Obiedkov, 2008], pages 258–272. 106

[Jay et al., 2006b] Jay, N., Napoli, A., and Kohler, F. (2006b). Cancer patient flows discovery in drg databases. *Stud Health Technol Inform*, 124 :725–730. 105

[Jay et al., 2006c] Jay, N., Napoli, A., and Kohler, F. (2006c). Mise en évidence de filières de soins dans le traitement du cancer à l'aide des motifs fréquents et treillis de galois. In *XIX Journnées EMOIS*. 105

[Jay et al., 2007] Jay, N., Napoli, A., and Kohler, F. (2007). Mise en évidence de filières de soins dans le traitement du cancer à l'aide des motifs fréquents et treillis de galois. In *12èmes Journées Francophones d'Informatique Médicale, Bamako*. 105

[Jenicek and Stachenko, 2003] Jenicek, M. and Stachenko, S. (2003). Evidence-based public health, community medicine, preventive care. *Med Sci Monit*, 9(2) :SR1–SR7. 62

[Kipke and Wille, 1987] Kipke, U. and Wille, R. (1987). Formale begriffsanalyse erläutert an einem wortfeld. *LDV–Forum*, 5. 34, 35

[Kuznetsov, 1989] Kuznetsov, S. (1989). Interpretation on graphs and complexity characteristics of a search for specific patterns. *Automatic Documentation and Mathematical Linguistics*, 24(1) :37–45. 23

[Kuznetsov et al., 2007] Kuznetsov, S., Obiedkov, S., and Roth, C. (2007). Reducing the representation complexity of lattice-based taxonomies. In Priss, U., Polovina, S., and Hill, R., editors, *Proc. of ICCS 15th Intl Conf Conceptual Structures*, volume 4604 of *LNCS/LNAI*, pages 241–254. Springer. 52, 55, 56, 78, 79

[Kuznetsov, 1990] Kuznetsov, S. O. (1990). Stability as an estimate of the degree of substantiation of hypotheses derived on the basis of operational similarity. *Nauchn. Tekh. Inf., Ser.2 (Automat. Document. Math. Linguist.)*, 12 :21–29. 52

[Kuznetsov, 2007] Kuznetsov, S. O. (2007). On stability of a formal concept. *Annals of Mathematics and Artificial Intelligence*, 49 :101–115. 52, 55

[Kuznetsov and Obiedkov, 2001] Kuznetsov, S. O. and Obiedkov, S. A. (2001). Algorithms for the construction of concept lattices and their diagram graphs. In *Principles of Data Mining and Knowledge Discovery : 5th European Conference, PKDD 2001*, pages 289–300, Freiburg, Germany. 23, 24

[Lerman, 1970] Lerman, I. (1970). *Les bases de la classification automatique.* Gauthier-Villars Éditeurs, Paris. 57

[Lindig, 1995] Lindig, C. (1995). Concept-based component retrieval. working notes of the IJCAI-95 workshop on formal approaches to the reuse of plans, Proofs. 36

[Medina and Obiedkov, 2008] Medina, R. and Obiedkov, S. A., editors (2008). *Formal Concept Analysis, 6th International Conference, ICFCA 2008, Montreal, Canada, February 25-28, 2008, Proceedings*, volume 4933 of *Lecture Notes in Computer Science*. Springer. 118, 120

[Messai et al., 2007] Messai, N., Devignes, M.-D., Napoli, A., and Smaïl-Tabbone, M. (2007). Traitement d'attributs inter-dépendants pour la recherche d'information par treillis. In Trichet, F. and Fiorino, H., editors, *Ingénierie des connaissances (IC-2007), Grenoble*, pages 109–120. Cépadues éditions, Toulouse. 34

[Mineau and Godin, 1995] Mineau, G. and Godin, R. (1995). Automatic structuring of knowledge bases by conceptual clustering. *Knowledge and Data Engineering, Transactions on*, 7(5) :824–829. 31

[Naiditch, 1993] Naiditch, M. (1993). Contact, épisode, réseau et filière : quelle(s) porte(s) d'entrée et quelle(s) place(s) pour les systèmes de classification ? In *Informatique Médicale et Stratégies Hospitalières*, volume 6 of *Informatique et Santé*, pages 133–44. Springer-Verlag France. 68

[Naiditch, 2000] Naiditch, M. (2000). Modélisation des trajectoires : problèmes méthodologiques. *ITBM-RBM*, 21(5) :307–12. 68

[Norris, 1978] Norris, E. (1978). An algorithm for computing the maximal rectangles in a binary relation. *Revue Roumaine de Mathématiques Pures et Appliquées*, 23(2) :243–250. 23

[Nourine and Raynaud, 1999] Nourine, L. and Raynaud, O. (1999). A fast algorithm for building lattices. *Inf. Process. Lett.*, 71(5-6) :199–204. 23

[Pasquier et al., 1999a] Pasquier, N., Bastide, Y., Taouil, R., and Lakhal, L. (1999a). Closed set based discovery of small covers for association rules. In *Proc. 15èmes Journées Bases de Données Avancées, BDA*, pages 361–381. 46

[Pasquier et al., 1999b] Pasquier, N., Bastide, Y., Taouil, R., and Lakhal, L. (1999b). Discovering frequent closed itemsets for association rules. In Beeri, C. and Buneman, P., editors, *ICDT*, volume 1540 of *Lecture Notes in Computer Science*, pages 398–416. Springer. 46

[Pasquier et al., 1999c] Pasquier, N., Bastide, Y., Taouil, R., and Lakhal, L. (1999c). Efficient mining of association rules using closed itemset lattices. *Journal of Information Systems*, 24 :25–46. 30, 31, 45, 46

[Pei et al., 2000] Pei, J., Han, J., and Mao, R. (2000). Closet : An efficient algorithm for mining frequent closed itemsets. In *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 21–30. 46

[Pogel and Ozonoff, 2008] Pogel, A. and Ozonoff, D. (2008). Contingency structures and concept analysis. In [Medina and Obiedkov, 2008], pages 305–320. 110

[Priss, 2005] Priss, U. (2005). Linguistic applications of formal concept analysis. In [Ganter et al., 2005], pages 149–160. 34

[R Development Core Team, 2007] R Development Core Team (2007). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. 98

[Richard and Richard, 1995] Richard, J.-F. and Richard, A. (1995). *Cours de psychologie*, chapter Les bases des fonctionnements cognitifs, page 431. Dunod, Paris. 28

[Riou and Jarno, 2000] Riou, F. and Jarno, P. (2000). Représentation et modélisation des trajectoires de soins. *ITBM-RBM*, 21(5) :313–17. 67, 68

[Roth, 2006] Roth, C. (2006). Co-evolution in epistemic networks – reconstructing social complex systems. *Structure and Dynamics : eJournal of Anthropological and Related Sciences*, 1(3). 56, 78

[Roth et al., 2006] Roth, C., Obiedkov, S., and Kourie, D. G. (2006). Towards concise representation for taxonomies of epistemic communities. In Yahia, S. B. and Nguifo, E. M., editors, *CLA 4th International Conference on Concept Lattices and their Applications*, pages 205–218, Tunis. Faculté des Sciences de Tunis. 56, 78, 79

[Schmitz et al., 2002] Schmitz, C., Staab, S., Studer, R., Stumme, G., and Tan, J. (2002). Accessing distributed learning repositories through a courseware watchdog. In *Proceedings of the ECAI 2002 Workshop on Machine Learning and Natural Language Processing for Ontology Engineering (OLT 2002)*. 33

[Seno and Karypis, 2002] Seno, M. and Karypis, G. (2002). Slpminer : an algorithm for finding frequent sequential patterns using length-decreasing support constraint. In *Data Mining, 2002. ICDM 2002. Proceedings. 2002 IEEE International Conference on*, pages 418–425. 100

[Snelting and Tip, 1998] Snelting, G. and Tip, F. (1998). Reengineering class hierarchies using concept analysis. In *SIGSOFT '98/FSE-6 : Proceedings of the 6th ACM SIGSOFT international symposium on Foundations of software engineering*, pages 99–110, New York, NY, USA. ACM. 36

[Snelting and Tip, 2000] Snelting, G. and Tip, F. (2000). Understanding class hierarchies using concept analysis. *ACM Trans. Program. Lang. Syst.*, 22(3) :540–582. 36

[Sokal, 1968] Sokal, R. (1968). Phenetic taxonomy : Theory and methods. *Annual Review of Ecology and Systematics*, 17 :423–442. 57

[Stefanelli, 2001] Stefanelli, M. (2001). The socio-organizational age of artificial intelligence in medicine. *Artif Intell Med*, 23(1) :25–47. 63, 64, 65

[Stumme, 2002a] Stumme, G. (2002a). Efficient data mining based on formal concept analysis. In *Lecture Notes in Computer Science*, volume 2453, page 534. Springer. 31, 47

[Stumme, 2002b] Stumme, G. (2002b). Formal concept analysis on its way from mathematics to computer science. In Priss, U., Corbett, D., and Angelova, G., editors, *ICCS*, volume 2393 of *Lecture Notes in Computer Science*, pages 2–19. Springer. 30

[Stumme and Maedche, 2001] Stumme, G. and Maedche, A. (2001). Fca-merge : Bottom-up merging of ontologies. In Nebel, B., editor, *IJCAI*, pages 225–234. Morgan Kaufmann. 32

[Stumme et al., 2001] Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., and Lakhal, L. (2001). Intelligent structuring and reducing of association rules with formal concept analysis. In Baader, F., Brewka, G., and Eiter, T., editors, *KI/ÖGAI*, volume 2174 of *Lecture Notes in Computer Science*, pages 335–350. Springer. 45

[Stumme et al., 2002] Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., and Lakhal, L. (2002). Computing iceberg concept lattices with titanic. *Data Knowl. Eng.*, 42(2) :189–222. 30, 45, 46, 48, 78

[Szathmary, 2006] Szathmary, L. (2006). *Méthodes symboliques de fouille de données avec la plate-forme Coron*. PhD Thesis in Computer Sciences, Université Henri Poincaré – Nancy 1, France. (title in English : Symbolic Data Mining Methods with the Coron Platform). 26

[Szathmary et al., 2007] Szathmary, L., Napoli, A., and Valtchev, P. (2007). Towards rare itemset mining. In *Proceedings of the IEEE International Conference on Tools with Artificial Intelligence (ICTAI), Patras, Greece*. IEEE Computer Society. 56, 110

[Tilley et al., 2005] Tilley, T., Cole, R., 0002, P. B., and Eklund, P. W. (2005). A survey of formal concept analysis support for software engineering activities. In [Ganter et al., 2005], pages 250–271. 36, 37

[Trombert-Paviot et al., 2000] Trombert-Paviot, B., Rodrigues, J. M., Rogers, J. E., Baud, R., van der Haring, E., Rassinoux, A. M., Abrial, V., Clavel, L., and Idir, H. (2000). Galen : a third generation terminology tool to support a multipurpose national coding system for surgical procedures. *Int J Med Inform*, 58-59 :71–85. 111

[Trétarre et al., 2004] Trétarre, B., Guizard, A.-V., Fontaine, D., and les membres du réseau Francim et le CépiDc-Inserm (2004). Cancer du sein chez la femme : incidence et mortalité, france 2000. *Bulletin Épidémiologique Hebdomadaire*, 44 :209–210. 96

[Valtchev et al., 2003] Valtchev, P., Grosser, D., Roume, C., and Hacene, M. (2003). Galicia : an open platform for lattices. in using conceptual structures. In *Contributions to the 11th Intl. Conference on Conceptual Structures (ICCS'03)*, page 241–254. Shaker Verlag. 24, 80

[Valtchev et al., 2004] Valtchev, P., Missaoui, R., and Godin, R. (2004). Formal concept analysis for knowledge discovery and data mining : The new challenges. In [Eklund, 2004], pages 352–371. 30

[Valtchev et al., 2002] Valtchev, P., Missaoui, R., and Lebrun, P. (2002). A partition-based approach towards constructing galois (concept) lattices. *Discrete Math.*, 256(3) :801–829. 23

[Vogt and Wille, 1994] Vogt, F. and Wille, R. (1994). Toscana - a graphical tool for analyzing and exploring data. In Tamassia, R. and Tollis, I. G., editors, *Graph Drawing*, volume 894 of *Lecture Notes in Computer Science*, pages 226–233. Springer. 24

[Vogt et al., 1991] Vogt, F. C., Wachter, and Wille, R. (1991). *Classification, Data Analysis and Knowledge Organization*, chapter Data Analysis based on a Conceptual File Hans-Hermann Bock„ pages 131–140. Springer Verlag, Berlin. 21

[Wang et al., 2003] Wang, J., Han, J., and Pei, J. (2003). Closet+ : searching for the best strategies for mining frequent closed itemsets. In Getoor, L., Senator, T. E., Domingos, P., and Faloutsos, C., editors, *KDD*, pages 236–245. ACM. 46

[Weiner et al., 1991] Weiner, J., Starfield, B., Steinwachs, D., and Mumford, L. (1991). Development and application of a population-oriented measure of ambulatory care case mix. *Medical Care,*; 29(5) : 452-72.(5) :452–72. 69

[Wille, 1982] Wille, R. (1982). Restructuring lattice theory : an approach based on hierarchies of concepts. In Rival, I., editor, *Ordered Sets*. Reidel. 7, 13, 30

[Wille, 1989] Wille, R. (1989). *Algorithms and order*, chapter Lattices in data analyse : how to draw them with a computer, pages 33–58. Kluwer, Dordrecht-Boston. 21

[Wille, 1999] Wille, R. (1999). Conceptual landscapes of knowledge : a pragmatic paradigm for knowledge processing. In Gaul, W. and Locarek-Junge, editors, *Classification in the Information Age*, pages 344–356. Springer, Heidelberg. 29

[Wille, 2000] Wille, R. (2000). Boolean concept logic. In Ganter, B. and Mineau, G. W., editors, *ICCS*, volume 1867 of *Lecture Notes in Computer Science*, pages 317–331. Springer. 29

[Wille, 2005] Wille, R. (2005). Formal concept analysis as mathematical theory of concepts and concept hierarchies. In [Ganter et al., 2005], pages 1–33. 28

[Yan et al., 2003] Yan, X., Han, J., and Afshar, R. (2003). Clospan : Mining closed sequential patterns in large databases. In Barbará, D. and Kamath, C., editors, *SDM*. SIAM. 48

[Yevtushenko, 2000] Yevtushenko, S. (2000). System of data analysis "concept explorer". In *Proceedings of the 7th national conference on Artificial Intelligence KII-2000*, pages 127–134, Russia. (In Russian). 24

[Zabezhailo et al., 1987] Zabezhailo, M., Ivashko, V., Kuznetsov, S., Mikheenkova, M.A. nad Khazanovskii, K., and Anshakov, O. (1987). Algorithms and programs of the jsm-method of automatic hypothesis generation. *Automatic Documentation and Mathematical Linguistics*, 21 :1–14. 23

[Zaki and Hsiao, 2002] Zaki, M. J. and Hsiao, C.-J. (2002). Charm : An efficient algorithm for closed itemset mining. In Grossman, R. L., Han, J., Kumar, V., Mannila, H., and Motwani, R., editors, *SDM*. SIAM. 30, 45, 46

[Zaki and Hsiao, 2005] Zaki, M. J. and Hsiao, C.-J. (2005). Efficient algorithms for mining closed itemsets and their lattice structure. *IEEE Trans. Knowl. Data Eng.*, 17(4) :462–478. 46

[Zaki et al., 2005] Zaki, M. J., Parimi, N., De, N., Gao, F., Phoophakdee, B., Urban, J., Chaoji, V., Hasan, M. A., and Salem, S. (2005). Towards generic pattern mining. In Ganter, B. and Godin, R., editors, *ICFCA*, volume 3403 of *Lecture Notes in Computer Science*, pages 1–20. Springer. 110

Enjeu majeur de santé publique, les maladies chroniques nécessitent souvent une approche multidisciplinaire et des contacts multiples entre le patient et le système de soins. La maîtrise de ce parcours, la trajectoire de soins, est un gage de qualité des soins, qualité de vie et d'efficience médico-économique. Pour maîtriser les trajectoires de soins, il faut les connaître. Or, à ce jour en France, aucun système d'information en santé à grande échelle n'est conçu pour décrire les trajectoires de soins, encore moins d'en établir une typologie. Malgré tout, d'énormes quantités de données sont produites chaque année par le système de soins.

Dans ce travail, nous proposons un système d'analyse et de représentation des trajectoires de soins à partir de données récoltées à l'origine pour d'autres utilisations. Ce système est fondé sur l'Analyse Formelle de Concepts (AFC), une méthode de classification conceptuelle capable de découvrir des liens naturels dans les données d'un tableau binaire et de les représenter sous forme de treillis de concepts. Nous montrons les apports de l'AFC dans la compréhension du fonctionnement du système de soins et les perspectives en termes d'exploration des réseaux sociaux en général.

Par ailleurs, nous étudions, combinons et comparons deux mesures d'intérêt pour réduire la complexité des grands treillis et sélectionner les connaissances les plus pertinentes : la stabilité et le support d'un concept. Dans un parallèle avec la recherche de motifs fréquents simple et séquentiels, nous proposons une méthode de classification non supervisée des trajectoires de soins qui présente d'intéressantes capacités de visualisation et d'interprétabilité.

Mots-clés : Analyse formelle de concepts, treillis, extraction des connaissances, classification, trajectoire de soins, stabilité.

Abstract

Chronic diseases are a major concern in public health. They require a multidisciplinary approach and chronic patients have multiple contacts with the healthcare system. In order to improve quality of life, quality and efficiency of healthcare, the care process of chronic patients must be assessed. This implies that the whole succession of health events of such patients should be monitored. From now on, in France, no information system was design to describe cross-institutional and long-term patient care trajectories. Moreover, there have been few attempts to classify and build a typology of patient trajectories. However, enormous amount of data are collected each year by several healthcare information systems.

In this work, we propose a new approach to analyze and represent patient care trajectories from data that are originally collected for a different purpose. Our system is based on Formal Concept Analysis (FCA). FCA is a method of conceptual classification aiming at discovering natural relationships in a binary table and representing them in a concept lattice. We show the advantages of FCA for describing and understanding processes in the healthcare system and we give some perspectives for the exploration of social networks in general.

We also study and compare two measures of interestingness of concepts : stability and support. The measures are computed in order to reduce the complexity of big concept lattices and filter the most relevant knowledge. Drawing a parallel with itemset mining and sequential itemset mining, we propose an unsupervised classification method with interesting properties for visualization and interpretation.

Keywords : Formal concept analysis, lattice, knowlege discovery, classification, care trajectory, stability.