MATHÉMATIQUES ET SCIENCES HUMAINES

MARIE-CATHERINE DANIEL-VATONNE
COLIN DE LA HIGUERA

Les termes : un modèle algébrique de représentation et de structuration de données symboliques

Mathématiques et sciences humaines, tome 122 (1993), p. 41-63
<http://www.numdam.org/item?id=MSH_1993__122__41_0>

Math. Inf. Sci. hum. (30e année, n° 122, *1993*, p. 41-63)

LES TERMES : UN MODÈLE ALGÉBRIQUE DE REPRÉSENTATION ET DE STRUCTURATION DE DONNÉES SYMBOLIQUES

Marie-Catherine DANIEL-VATONNE[1], Colin DE LA HIGUERA[1]

RÉSUMÉ — *Nos travaux se situent dans le cadre de l'analyse conceptuelle des données. Notre objectif est de généraliser les représentations par variables binaires ou nominales en y adjoignant la modélisation de structures internes. Le problème est de ne pas perdre en complexité algorithmique ce qui est gagné en puissance de représentation. Selon ces considérations, décrire les données et des classes de données par des structures arborescentes semble un bon compromis. Le système de représentation que nous proposons s'appuie sur un modèle algébrique : les magmas. Il permet de construire des termes assimilables à des arborescences finies, étiquetées et typées. Leur interprétation est intuitive et ils autorisent les descriptions récursives. Une relation d'ordre naturelle, la généralisation, induit un treillis sur les termes. La construction des termes, leur comparaison dans l'ordre, le calcul des bornes supérieures et inférieures ont une complexité polynomiale. Ce modèle inclut le cadre binaire tout en conservant certaines propriétés. En particulier, nous montrons que l'on peut construire un treillis de Galois mettant en correspondance des ensembles d'objets et leurs descriptions par des termes. Une application est donnée à titre d'illustration, portant sur la transmission héréditaire du daltonisme.*

SUMMARY — The terms : an algebraic model of representation and structuration of symbolic data.
Our framework is concept analysis. Our goal is to generalize systems based on descriptions by nominal or binary variables, by taking into account the internal structure of data. The problem nevertheless is not to lose in algorithmic complexity what is gained in quality of the description. Under these considerations, ordered trees seem a good compromise. The representation system we propose is based on an algebraic model : magmas. Typed terms (isomorphic to oriented trees) are used to describe the data (and class characterization). Their interpretation is intuitive and recursive descriptions are allowed. A natural partial order, generalization, induces a lattice structure on these terms. Term construction, term comparison, computation of the supremum and infimum of sets of terms are all polynomially tractable problems. This model preserves most properties of the description by binary variables ; in particular we show how the Galois lattice between sets of objects and their description can be constructed. As an illustration this lattice is constructed from data related to the transmission of colour-blindness.

1. INTRODUCTION

Nos travaux se situent dans le cadre de l'analyse de données symboliques et structurées. Nous entendons par là des données décrites à partir d'un ensemble de symboles sur lequel on peut définir des structures discrètes. Notre approche est de type "conceptuelle" [7]. Si, par exemple, on se place dans le cadre de la classification automatique, l'analyse conceptuelle consiste à construire des classes qui ont de bonnes propriétés et dont la caractérisation est exprimée de

[1] Département Informatique Fondamentale, LIRMM, Montpellier.

manière symbolique, dans le langage de représentation des données. Ce sont ces caractérisations que l'on qualifie de concepts. De la même manière, en analyse discriminante, l'analyse conceptuelle consiste à construire des règles, ou systèmes de règles, exprimées sous forme symbolique et ayant un bon pouvoir de séparation des classes.

Notre propos est l'exposé d'un langage permettant de représenter à la fois des données et des caractérisations de classe.

Nous nous intéressons plus particulièrement aux données dont la représentation peut d'une part inclure des descriptions apparentées à des variables binaires ou nominales et d'autre part admettre la description de structures internes. Par exemple, des piles de cubes et de boules colorés se modélisent à partir d'attributs : *couleur*, *forme*, mais aussi d'une relation entre objets : *un cube vert **est sur** une boule jaune*.

Des travaux anciens répondent à ce type de préoccupation. Winston [16] utilise des réseaux sémantiques, proches de graphes étiquetés. Dietterich et Michalski [5] gèrent des données assimilables à des conjonctions de la logique des prédicats. Les domaines réels modélisables sont moins restreints que si seuls les attributs sont autorisés mais la plupart des algorithmes de traitement sont NP-complets [11]. En effet une formule logique caractérise un ensemble de données si elle est impliquée par elles. Ceci fait appel à un algorithme, NP-complet dans le cas général. De même, un ensemble de graphes est caractérisé par un sous-graphe qui leur est commun et l'isomorphisme de sous-graphe n'est polynomial que dans certains cas.

Notre approche à l'instar de Gascuel [6], Liquière [12] tient compte de l'équilibre à trouver entre les types potentiels de connaissances réelles représentables et la complexité des algorithmes de traitement. Le modèle choisi s'appuie sur celui des magmas utilisés jusqu'à présent dans le cadre de la sémantique algébrique de programmes ([8], [10]). Les données sont décrites par des termes qui sont des structures arborescentes non-commutatives, typées, étiquetées et finies. Le vocable "terme" est choisi par référence au mode de génération qui est très proche de celui des mots d'une grammaire et encore plus de la construction syntaxique des termes de la logique du premier ordre.

Notre choix est dicté par la volonté d'obtenir un système informatique efficace. Une opération de base d'un système d'analyse conceptuelle est de comparer ou *apparier* deux descriptions, c'est-à-dire déterminer leur(s) plus grande(s) partie(s) commune(s). L'appariement d'arborescences est polynomial, mais cette famille de graphes semble être "à la limite". L'appariement de forêts, par exemple, est NP-complet (la vérification qu'une forêt est sous-graphe d'un arbre étant déjà NP-complet).

Une autre raison de ce choix est que l'interprétation d'un terme est directe. Intuitivement un sommet peut être considéré :

- comme une fonction typée, dont les arguments sont typés et en nombre fixe. Par exemple *"un daltonien conduit une jeep"* se traduit par une fonction *conduire* de type *action* prenant un premier argument de type *humain* et un second de type *véhicule*.

- comme un objet constitué de plusieurs composantes. Par exemple, une *voiture* est composée d'une *carrosserie*, d'un *moteur*, de *4 roues*...

- comme un questionnaire permettant de donner des caractéristiques (ou attributs) de manière organisée. Ainsi, l'un des successeurs du sommet *voiture* pourra être *quatre roues motrices*. A l'inverse, cet attribut ne figurera pas parmi les descendants possibles de *moto*.

Un terme peut aussi décrire partiellement une donnée. On peut ainsi ne pas spécifier que c'est un daltonien qui conduit une jeep. L'interprétation du terme correspondant est *"humain conduisant une jeep"*. Il devient alors une description partielle d'autres données : *"un daltonien conduit une jeep"*, *"une astronaute conduit une jeep"*, *"Charlemagne conduit une jeep"*... Autrement dit, il représente la classe des données qu'il décrit.

Enfin, ce modèle permet de décrire des données ou des classes de données de manière récursive sur le mode : *humain dont le père est un humain dont le père ...*

Dans la partie II, le modèle proposé est décrit de manière détaillée. Un rapport est établi entre une modélisation par variables nominales et une modélisation par termes. La partie III met en évidence l'existence d'une forme normale pour chaque terme. Dans la partie IV, nous montrons que la relation d'ordre naturelle sur les termes, la "généralisation" ou "inclusion", induit une structure de treillis. Toute tâche d'analyse conceptuelle se ramène alors peu ou prou à une exploration de ce treillis. A titre d'illustration, nous montrons dans la partie V qu'il existe une correspondance de Galois entre le treillis des termes et le treillis des parties d'un ensemble d'objets. Cette correspondance est exploitée, comme dans le cadre binaire (Barbut & Monjardet [1], Wille [15]) pour construire le treillis de Galois, ou treillis de concepts, liant un ensemble d'objets et leurs descriptions. Un exemple d'application portant sur la transmission héréditaire du daltonisme est présenté.

Avant de présenter notre modèle, précisons que notre démarche privilégie un aspect algorithmique, au sens où les définitions sont généralement récursives et aisément transformables en algorithmes. Nous commençons par les présenter puis nous montrons qu'elles coincident avec les notions usuelles auxquelles elles font référence.

II. REPRÉSENTATION DES DONNÉES

Le modèle des magmas ([8], [10]) permet de construire et manipuler des structures finies, non commutatives, étiquetées et typées appelés *termes*, de donner à ceux-ci une interprétation formelle en incluant la notion de système de réécriture. Comme nous nous intéressons essentiellement dans cet article à la représentation de données par des termes, nous n'abordons pas le niveau de la sémantique formelle. C'est pourquoi, dans un souci de clarté, nous exposons une version simplifiée du modèle et préférons parler d'espaces de termes plutôt que de magma.

Une première section définit la signature et les espaces de termes qu'elle définit en intention et permet de générer. En second lieu, nous nous intéressons aux termes décrivant partiellement des données, ils représentent des classes de données. Une troisième section définit une relation de généralisation entre termes analogue à une relation d'inclusion entre classes de données. Enfin, des rapports entre représentation par variables nominales et représentation par termes sont établis dans une dernière section.

II.1. Description des données par des termes

Nous définissons en premier lieu la signature qui est la base de tout le modèle. Elle permet à la fois l'existence et la construction des termes.
On note $[n]$ l'ensemble $\{1,2,..,n\}$ pour n entier, avec $[0] = \varnothing$.

DÉFINITION : Signature

Une *signature* est un quintuplet $(S, F, \sigma, \alpha, b)$ où :

- S est un ensemble fini de *types*
- F est un ensemble fini de *symboles*
- σ est une application de F dans S. Pour $f \in F$, $\sigma(f)$ est le *type* de f.
- α est une application de F dans S^* (S^* est l'ensemble des mots formés à partir des éléments de S). Pour $f \in F$, $\alpha(f)$ fixe l'ordre et le type des *arguments* de f. (Quand f n'a pas d'arguments, $\alpha(f)$ est le mot vide noté ε)
- $b \in S$ est appelé *type de base.*

Une signature permet de générer autant d'espaces de termes qu'il y a de types. Voici, à travers leur définition, le mode de construction de ces espaces :

DÉFINITION : Espace de termes

Soit une signature $SI = (S, F, \alpha, \sigma, b)$

On définit pour chaque type s de S un *espace de termes* $\mathcal{T}_s$ tel que :

$\mathcal{T}_s{}^0 = \{ f \in F , \sigma(f) = s , \alpha(f) = \varepsilon \}$

$\mathcal{T}_s{}^{k+1} = \mathcal{T}_s{}^k \cup \{ f(t_1,..,t_n)$ avec $f \in F$, $\sigma(f) = s$, $\alpha(f) = s_1..s_n$ et $\forall\ i \in [n]\ t_i \in \mathcal{T}_{s_i}{}^k \}$

$\mathcal{T}_s = \bigcup_{k \in \mathbf{N}} \mathcal{T}_s{}^k$

Nous profitons de cette définition pour préciser que le plus petit indice k tel que $t \in \mathcal{T}_s{}^k$ est appelé *hauteur* de t et est noté $h(t)$. Nous y faisons peu référence si ce n'est que la plupart des propriétés sur les termes peuvent se démontrer par récurrence sur cette hauteur.

Remarquons cependant que la hauteur est finie et qu'il découle des deux définitions précédentes que chaque terme contient un nombre fini de symboles. Par contre, $\mathcal{T}_s$ peut, lui, être infini car la hauteur des termes n'est pas bornée.

Les données à modéliser sont de même nature. C'est celle-ci qui est représentée par le type de base. L'espace $\mathcal{T}_b$ est donc l'espace des termes décrivant les données potentielles.

Exemple 1

On désire modéliser des piles de cubes et de boules de couleur verte ou jaune. La signature est :
$S = \{p, f, c\}$ (p pour pile, f : forme, c : couleur) $F = \{pile, finpile, boule, cube, jaune, vert\}$
σ et α sont données dans la table suivante

F	σ	α
pile	p	*fcp*
finpile	p	ε
boule	f	ε
cube	f	ε
jaune	c	ε
vert	c	ε

p est le type de base et les piles sont décrites par des termes appartenant à $\mathcal{T}_p$.
Ainsi une boule verte sur un cube jaune s'écrit : ***pile (cube, jaune, pile (boule, vert, finpile))***

Une représentation graphique naturelle d'un terme consiste à lui associer une arborescence dont les sommets sont étiquetés par des symboles et dont les successeurs directs de chaque sommet sont totalement ordonnés (cf figure 1).

C'est pourquoi nous empruntons le vocabulaire des arborescences :
Soit $t = f(t_1,..,t_n)$ un terme,

Outre la hauteur définie plus haut, nous appelons :

- *sommet* chaque occurence d'un symbole de t

- f est ***racine*** de t (et donc $\mathcal{T}_s$ est l'espace des termes dont la racine est de type s)

- si f n'a pas d'arguments, il est *feuille* de t

- $t_1,..,t_n$ sont *sous-termes* de t

Classiquement la représentation d'un terme place la racine en haut du graphique, les successeurs d'un sommet se lisent de gauche à droite.

Exemple 2

Le terme de l'exemple précédent *pile (cube, jaune, pile (boule, vert, finpile))* se représente par :

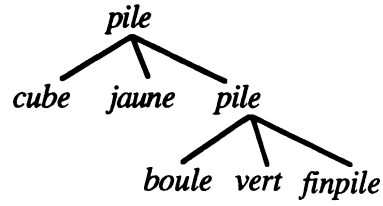

Figure 1

II.2. Description partielle

Pour représenter une information dont la teneur peut-être considérée comme inconnue ou inintéressante, nous utilisons un symbole particulier. Son type est la nature de l'information et il ne possède pas d'arguments puisque ceux-ci représenteraient la teneur de l'information.

On peut ainsi faire correspondre à chaque type s, un symbole Ω_s interprétable par : "*il existe une information de type s mais on ne connaît pas ou on ne s'intéresse pas à sa valeur*".

Il suffit pour cela de considérer les signatures pour lesquelles on distingue dans l'ensemble des symboles F, un ensemble $I = \{\Omega_s / s \in S , \sigma(\Omega_s) = s \text{ et } \alpha(\Omega_s) = \varepsilon\}$. Les termes contenant au moins un des symboles de I sont des descriptions partielles relativement aux termes dont tous les symboles ($f \notin I$) représentent une information spécifiée.

C'est à partir de cette notion de description partielle que nous allons générer un espace d'analyse permettant de comparer, regrouper, classer..., des descriptions de données. Cet espace d'analyse est pour nous un espace de termes issu d'une signature dont l'ensemble des symboles contient I. Nous considérons par la suite $\mathcal{T}_s$ comme issu d'une telle signature.

Exemple 3

On complète l'exemple précédant en ajoutant les symboles de I. La signature devient :
$S = \{p, f, c\}$ $F = \{pile, finpile, boule, cube, jaune, vert\} \cup \{\Omega_p, \Omega_f, \Omega_c\}$
σ et α sont données dans la table suivante :

F	σ	α	complétée par	I	σ	α
pile	p	fcp		Ω_p	p	ε
finpile	p	ε		Ω_f	f	ε
boule	f	ε		Ω_c	c	ε
cube	f	ε				
jaune	c	ε				
vert	c	ε				

Les piles sont décrites par des termes appartenant à $\mathcal{T}_p$. On peut ainsi représenter une pile dont la base est un cube par :

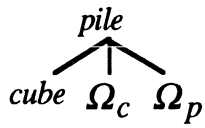

Figure 2

Pour simplifier (mais par abus de langage), un élément de I est désigné par Ω quand le type n'est pas utile ou n'a pas besoin d'être précisé.

II.3. Relation d'ordre

En se référant à l'exemple précédent "*une pile dont la base est un cube*" est une description partielle de toute pile dont la base est un cube. La relation entre termes qui traduit cette idée est appelée relation de généralisation : *t généralise t'*, notée $t \leq t'$, si t est une description partielle de t' .

DÉFINITION : Relation de généralisation

Soient $t \in \mathcal{T}_S$, et $t' \in \mathcal{T}_{S'}$

$t \leq t' \Leftrightarrow \sigma(t) = \sigma(t') = s$

et soit $t = \Omega_s$,

soit $t = f(t_1,..,t_n)$, $t' = f(t'_1,..,t'_n)$ et $\forall i \in [n]\ t_i \leq t'_i$.

PROPRIÉTÉS

• Ω_S est le plus petit élément de $\mathcal{T}_S$ (1)

La preuve est immédiate de par la définition.

• $\leq$ est une relation d'ordre partiel (2)

La preuve est immédiate de par la définition.

Exemple 4

$\mathcal{T}_p$ est l'espace des termes représentant toutes les piles possibles (selon la signature donnée dans l'exemple 3). La figure 3 donne une vue partielle de son graphe de couverture.

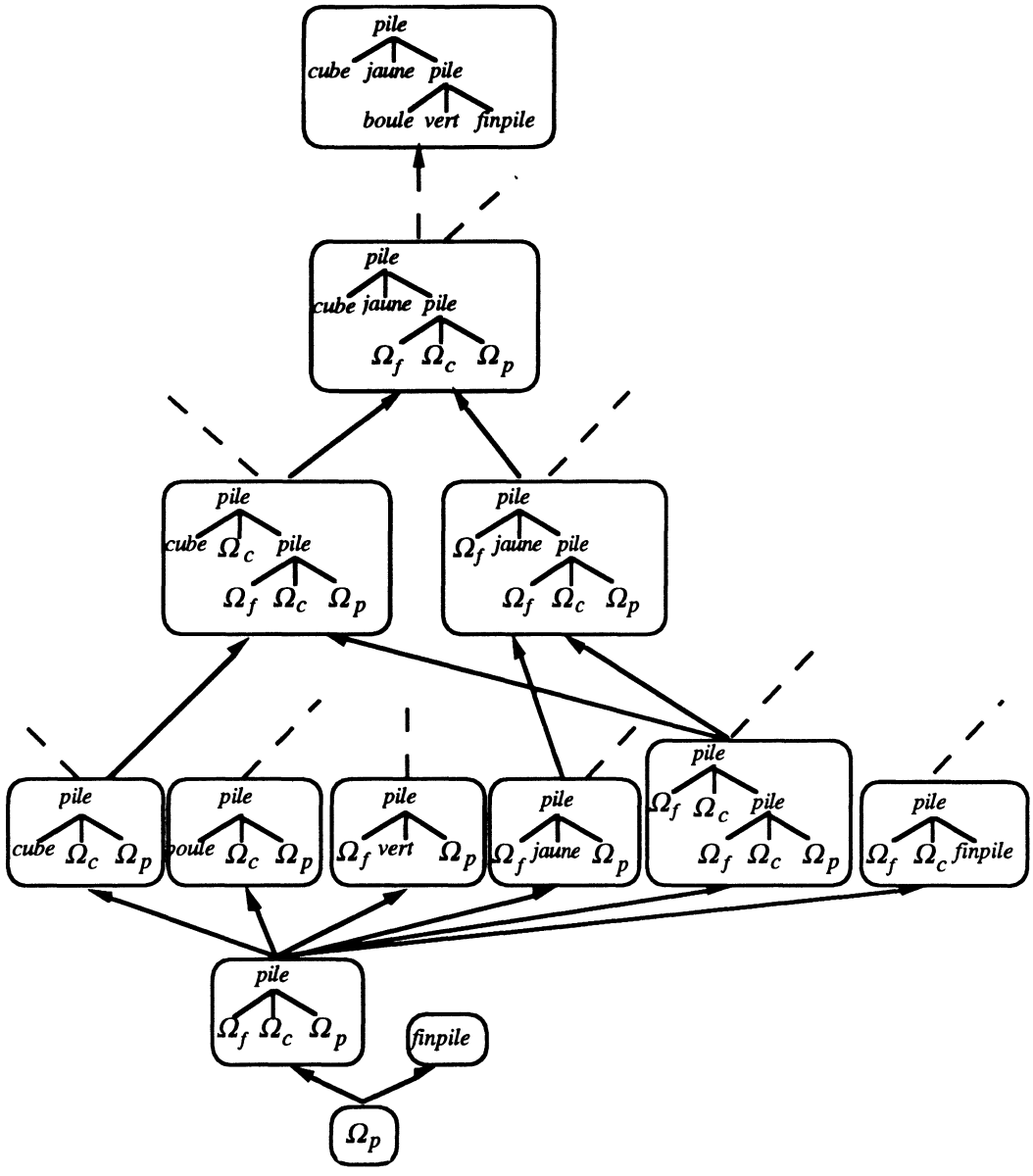

Figure 3

Intuitivement la construction de $\mathcal{T}_p$ consiste à trouver tous les successeurs directs d'un terme en remplaçant successivement chacun de ses Ω (s'il y en a) par chacun des plus petits sous-termes dont la racine est du même type que Ω, puis à itérer le processus sur les nouveaux termes construits.

Par exemple, Ω_p s'interprète par : *est une pile* et généralise directement *est une pile vide* (*finpile*) et *est une pile possédant un objet* (*pile* (Ω_f, Ω_c, Ω_p)).

II.4. Termes et variables nominales

Cette partie précise les rapports existants entre une représentation par variables nominales et une représentation par termes. Ces rapports ne sont pas véritablement formalisés mais nous pensons que les intuitions suggérées sont suffisamment convaincantes.

• *La représentation par variables nominales* est classiquement définie de la façon suivante :

Soit $V = \{V_1,..,V_n\}$ un ensemble fini de variables nominales.

A chaque variable V_i de V est associé un domaine fini $D_i \cup \{inconnue\}$ où
- D_i est un ensemble de valeurs possibles
- *inconnue* représente l'absence de valeur.

Une description v par variables nominales est une suite de valeurs $(v_1,..,v_n)$ avec :
$v_i \in D_i \cup \{inconnue\}$ et v_i représente la valeur (ou l'absence de valeur) de la variable V_i .

Exemple 5
$V = \{V_f, V_c\}$ (f pour forme, c pour couleur),
$D_f = \{cube, boule\}$, $D_c = \{jaune, vert\}$,
l'objet "*cube vert*" se décrit par (*cube, vert*)

• *Ce modèle est inclus dans celui des termes.* Intuitivement il suffit de définir la signature suivante :
$SI = (S, F, \sigma, \alpha, o)$ avec :
$S = V \cup \{o\}$,
$F = \{objet\} \cup \bigcup_{i \in [n]} D_i \cup \{\Omega V_i\}$,
$\forall v_i \in D_i$, $\sigma(v_i) = V_i$, $\alpha(v_i) = \varepsilon$ et $\sigma(objet) = o$, $\alpha(objet) = V_1...V_n$

Exemple 6
La modélisation par variables nominales de l'exemple précédent peut se transformer en la signature suivante :
$S = \{o, v_f, v_c\}$ $F = \{objet, boule, cube, jaune, vert, \Omega_f, \Omega_c\}$ et le type de base est o
σ et α sont données dans la table suivante

F	σ	α
objet	o	fc
boule	v_f	ε
cube	v_f	ε
Ω_f	v_f	ε
jaune	v_c	ε
vert	v_c	ε
Ω_c	v_c	ε

l'objet "*cube vert*" se décrit par *objet* (*cube, vert*)

Remarquons que l'absence de valeur "*inconnue*" est traduite dans la signature par un Ω associé à chaque type correspondant à une variable. Cela signifie qu'implicitement nous avons accordé à "*inconnue*" la même sémantique qu'à Ω. Or dans une représentation par variables nominales, outre le fait que "*inconnue*" peut signifier variable inconnue ou inintéressante, on lui permet souvent d'être interprétée par "sans objet". Par exemple, la valeur "*masculin*" d'une variable "*sexe*" rend invalide une variable "*nombre de grossesses*". L'interprétation "sans objet" n'est pas valide pour un Ω. Le modèle des termes évite cette notion en permettant la modélisation de rapports entre informations. Dans le cas de l'exemple, seul le symbole "*féminin*" de type "*sexe*" a pour argument un symbole de type "*nombre de grossesse*". Cependant, on peut pour traduire littéralement une représentation par variables nominales, introduire si besoin dans la signature un symbole "*sans-objet*V_i" pour chaque type V_i . Il est utilisé quand "*inconnue*" signifie que la variable est invalide.

• *Une signature n'est pas toujours modélisable par des variables* puisqu'on ne peut pas trouver pour toute signature un ensemble fini de variables nominales et d'ensembles de valeurs. En effet la modélisation par variables implique un ensemble fini de descriptions. Un espace de termes, par exemple celui des piles, peut être infini.

• *Pour un ensemble donné de termes* on peut cependant imaginer un ensemble fini de variables nominales. L'idée est de décomposer tous les termes de l'ensemble en "chemins" partant de la racine et aboutissant à une feuille puis d'associer à chaque "chemin" une variable nominale dont les valeurs sont les symboles de même type que la feuille.

Exemple 7

Pour décrire des piles de 1 à 3 objets on a
V = {*forme-premier-objet, couleur-premier-objet, forme-second-objet, couleur-second-objet, forme-troisième-objet, couleur-troisième-objet*}
avec pour domaine de valeurs associé à :
forme-premier-objet : *boule, cube, inconnue*
couleur-premier-objet : *jaune,vert, inconnue*
forme-second-objet et *forme-troisième-objet* : *jaune, vert, inconnue, n'existe-pas*
couleur-second-objet et *couleur-troisième-objet* : *jaune, vert, inconnue, n'existe-pas*

• *Il y a trois désavantages principaux* à cette transformation :

Le premier est que la structuration des termes est perdue. Ainsi, l'implication entre (*forme-second-objet, n'existe-pas*) et (*forme-troisième-objet, n'existe-pas*) n'est pas explicite contrairement à la description correspondante par terme (Si après le premier objet la pile est terminée le symbole *finpile* le précise).

Ceci illustre aussi le second désavantage qui est : beaucoup de variables peuvent être sans objet pour certaines des données à décrire.

Enfin, si on ajoute une donnée dans l'ensemble des données à décrire, par exemple une pile de 4 objets, il faut redéfinir l'ensemble des variables nominales et redécrire les données initiales.

La signature permet d'éviter ces trois inconvénients.

III. SYMBOLES MUETS ET FORMES NORMALES

Cette partie montre que dans une représentation par signature on peut être amené à définir des symboles dont l'existence répond à des intérêts différents. Nous distinguons ainsi, au niveau de l'interprétation, trois sortes de symboles :
- ceux dont la présence est significative, autrement dit les symboles qui permettent la différenciation sémantique des termes : *un cube rouge* est différent d'*un cube vert*, il contient plus d'informations qu'*un cube sans couleur déterminée,*
- ceux qui servent à rendre plus explicite la représentation des données (voir ci-dessous : *III.1 Exemple 8 : symboles visuels),*
- ceux qui permettent la génération de sous-termes mais dont la signification est sans intérêt (voir ci-dessous : *III.2 Exemple 9 : symboles dédiés à la construction*).

Ces deux dernières espèces de symboles sont appelées symboles muets au sens où ils n'ajoutent pas d'informations significatives lors de l'interprétation d'un terme. Nous montrons que leur présence peut donner lieu à l'existence de termes sémantiquement équivalents. En effet nous allons voir que l'utilité d'un symbole muet dans un terme est fonction des sous-termes

qu'il induit. Si ceux-ci n'apportent pas non plus d'informations significatives, une opération de normalisation permet de les supprimer et définit le plus petit terme de sémantique équivalente.

Cette opération est basée sur une reconnaissance syntaxique des symboles muets. (*III.3 Symboles muets et normalisation*).

III.1. Exemple 8 : symboles visuels

Une modélisation des piles, différente de celle proposée dans l'exemple 3 peut considérer une pile comme un objet sur lequel se dresse une pile. Un objet est défini par une forme et une couleur. Ce qui donne la signature suivante :

$S = \{p, f, c, o\}$ $F = \{pile, finpile, boule, cube, jaune, vert, objet\} \cup \{\Omega_p, \Omega_f, \Omega_C, \Omega_O\}$

σ et α sont données dans la table suivante :

F	σ	α	complétée par	I	σ	α
pile	p	op		Ω_p	p	ε
finpile	p	ε		Ω_O	o	ε
objet	o	fc		Ω_f	f	ε
boule	f	ε		Ω_C	c	ε
cube	f	ε				
jaune	c	ε				
vert	c	ε				

On peut ainsi représenter une pile dont l'objet de base est un cube jaune par :

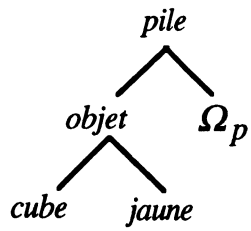

Figure 4

Cette représentation peut être considérée comme plus "parlante" que celle de l'exemple précédent mais elle n'ajoute pas d'informations et sémantiquement les deux termes de $\mathfrak{T}_O$ ci-dessous sont équivalents (*objet dont on ne connaît ni la forme, ni la couleur*).

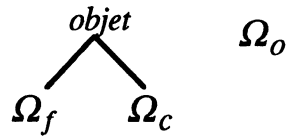

Figure 5

III.2. Exemple 9 : symboles dédiés à la construction

On considère des hommes daltoniens et on s'intéresse à la présence ou non du daltonisme chez leurs ascendants. Pour simplifier on ne représente que les ascendants paternels (un exemple complet est proposé dans la partie V).

La signature est la suivante :
S = {*homme, daltonisme*} (le type de base est *homme*)
F = {h, d, n} ∪ {Ω_h, Ω_d} (d pour daltonien et n pour non-daltonien)
σ et α sont données dans la table suivante :

F	σ	α	I	σ	α
h	*homme*	*daltonisme homme*	Ω_h	*homme*	ε
n	*daltonisme*	ε	Ω_d	*daltonisme*	ε
d	*daltonisme*	ε			

Ainsi un homme daltonien dont le père l'est aussi se représente par :

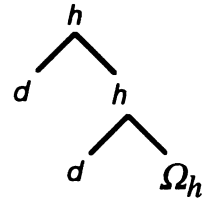

Figure 6

Le symbole h permet la construction récursive de l'ascendance. On ne lui accorde pas de signification (tout homme a un ascendant homme est trivial). C'est pourquoi le terme ci-dessous (*homme daltonien dont le père est daltonien et possédant un grand-père et un arrière-grand-père*) est équivalent au terme de la figure 6 ci-dessus.

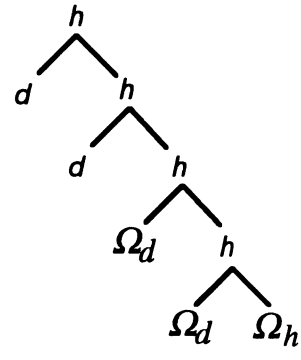

Figure 7

III.3. Symboles muets et normalisation

Cette section formalise les notions présentées dans les deux exemples précédents.

En premier lieu, on constate que ce qui différencie la nature sémantique du symbole *pile* de celle du symbole h est que le nombre d'occurences de *pile* indique le nombre d'objets. Ceci

grâce à *finpile* qui permet de terminer l'énumération et dont le sens est très différent de Ω_p. Il existe une contrepartie syntaxique à cette remarque sémantique : h est le seul symbole autre que Ω_h de type *homme* . De même *objet* (dont la raison d'être n'est que visuelle) est le seul symbole autre que Ω_o de type o. Nous appelons symboles muets, les symboles de F dont le type n'est partagé que par un Ω et nous définissons la forme normale d'un terme comme le terme de signification équivalente contenant un minimum de symboles. Ainsi la forme normale de *objet* (Ω_f, Ω_c) est Ω_o ; celle de *objet* (Ω_f, *jaune*) reste ce même terme.

DÉFINITION : Symbole muet
Un symbole $f \in F - I$ est **muet** si $\forall f' \in F$, $\sigma(f') = \sigma(f) \Rightarrow f = f'$

DÉFINITION : Forme normale d'un terme
Soit $t \in \mathcal{T}_S$
On définit la *forme normale* **de** t notée $N(t)$ par :
si $t = \Omega_S$ alors $N(t) = \Omega_S$
si $t = f$
  si f est muet alors $N(t) = \Omega_{\sigma(f)}$
  sinon $N(t) = f$
si $t = f(t_1,..,t_n)$
  si f est muet **et** $\forall i \in [n]$ $N(t_i) = \Omega$ alors $N(t) = \Omega_{\sigma(f)}$
  sinon $N(t) = f(N(t_1),.., N(t_n))$
Nous disons qu'un terme t est sous forme normale si $t = N(t)$.

PROPRIÉTÉS
$\forall t \in \mathcal{T}_S$,
• $N(t) \in \mathcal{T}_S$ (3)
• $N(t)$ est unique (4)

• La normalisation N considérée comme une application définit une *ouverture* [1] :
  • $N(t) = N(N(t))$ (idempotence) (5)
  • $N(t) \leq t$ (contraction) (6)
  • si $t \leq t'$ alors $N(t) \leq N(t')$ (monotonie croissante) (7)

PREUVES
• (3), (4), (5) et (6) sont des conséquences directes de la définition. En effet, celle-ci prend en compte tous les cas possibles de termes (existence de $N(t)$ pour tout terme t). Pour chacun d'eux, elle ne définit qu'une normalisation possible (unicité de $N(t)$) et assure l'idempotence. La forme normale appartient évidemment à $\mathcal{T}_S$. En se référant à la définition de $\leq$, il est évident que $N(t)$ généralise t .

•Prouvons (7) : si $t \leq t'$ alors $N(t) \leq N(t')$
Nous le faisons par récurrence sur la hauteur du terme t. Rappelons que la hauteur d'un terme est le plus petit indice n tel que $t \in \mathcal{T}_S{}^n$ et qu'on a par définition des espaces de termes $\forall k \in \mathbb{N}$, si $k \leq n$ alors $\mathcal{T}_S{}^k \subseteq \mathcal{T}_S{}^n$. Les définitions constructives que nous utilisons, donnent souvent lieu à des preuves par récurrence sur la hauteur des termes. Généralement, les deux premiers points d'une définition donnent les arguments pour un terme de hauteur 0, le troisième point donne le cas général et permet immédiatement de passer d'un terme de hauteur k à ses

sous-termes de hauteur strictement inférieure. Nous profitons de la propriété (7) pour développer une telle preuve.

- Si $h(t) = 0$, alors
  soit $t = \Omega_S$ et dans ce cas $N(t) = \Omega_S$ et $\forall\ t' \in \mathcal{T}_S$, $\Omega_S \leq N(t')$ (par (4) on a $N(t') \in \mathcal{T}_S$)
  soit $t = f$ et dans ce cas pour que $t \leq t'$ il faut que $t = t'$

- Supposons la proposition vraie pour $h(t) \leq k$ $(k \in \mathbb{N})$

- Si $h(t) = k+1$ alors $t = f(t_1,..,t_n)$
  Pour que $t \leq t'$ on a par définition de $\leq$, $t' = f(t'_1,..,t'_n)$ et $\forall\ i \in [n]$ $t_i \leq t'_i$

Si f est muet **et** $\forall\ i \in [n]$ $N(t_i) = \Omega$ alors $N(t) = \Omega_S$ et évidemment $\Omega_S \leq N(t')$
Sinon, $N(t) = f(N(t_1),..,N(t_n))$ et comme par définition des termes, $\forall\ i \in [n]$, $h(t_i) \leq k$ on est assuré par hypothèse de récurrence et par $t_i \leq t'_i$ que $N(t_i) \leq N(t'_i)$
on a donc bien $f(N(t_1),..,N(t_n)) \leq f(N(t'_1),..,N(t'_n))$ ◁

COROLLAIRE : $N(t)$ est le plus grand terme sous forme normale inférieur à t. (8)

PREUVE
En effet, pour tout terme sous forme normale, noté $N(t')$, si $N(t') \leq t$ alors $N(N(t')) \leq N(t)$ et comme $N(t')$ est sous forme normale $N(N(t')) = N(t')$ et donc $N(t') \leq N(t)$. ◁

Ceci termine la partie II définissant la représentation par termes. Avant d'explorer la structure d'un espace de termes induite par $\leq$, remarquons que la relation "*a la même forme normale*" est une relation d'équivalence et que $N(t)$ définit un représentant canonique de cette classe d'équivalence. Puisque t et $N(t)$ ont la même sémantique, nous ne conservons que les termes sous forme normale pour chaque espace $\mathcal{T}_S$: $\mathcal{N}_S = \{N(t) / t \in \mathcal{T}_S\}$. Nous réduisons ainsi l'espace des termes à explorer, en ne gardant que les termes de sémantiques différentes.

Nous travaillons désormais sur $\mathcal{N}_S$.

IV. STRUCTURE DE L'ESPACE DES TERMES

Nous abordons maintenant une étude de la structure générale de $\mathcal{N}_S$. Les définitions et propositions énoncées montrent que le modèle s'articule bien autour de notions simples en théorie des ordres et treillis ([1], [2]).

Nous montrons en premier lieu que $\mathcal{N}_S$ muni de $\leq$ est un inf-demi-treillis. Le calcul de la borne inférieure de deux termes consiste à rechercher la structure commune aux données qu'ils décrivent. Il peut être vu comme l'opération d'appariement dans un système d'analyse conceptuelle.

En second lieu nous complétons $\mathcal{N}_S$ pour qu'il devienne un treillis. Pour ce faire, nous introduisons la notion de compatibilité entre termes. Le calcul de la borne supérieure de deux termes en découle. Au niveau interprétation il consiste à rechercher l'ensemble des données décrites par les deux termes. Autrement dit l'intersection des deux ensembles de données représentés par les termes.

Pour la suite, nous posons par défaut : $t,t' \in \mathcal{T}_S$, $t = f(t_1,..,t_n)$ et $t' = f'(t'_1,..,t'_m)$. Si $n = 0$ (ou $m = 0$) cette notation reste correcte et $t = f$ (ou $t' = f'$).

IV.1. Inf-demi treillis

$\mathcal{T}_S$ est un inf-demi-treillis, si pour toute paire de termes on peut trouver une borne inférieure et s'il existe un minorant universel.
L'intuition de la borne inférieure est la suivante : le plus petit terme généralisant deux termes contient tout ce qui est commun aux deux termes et les informations différentes sont substituées par des Ω. On parle de *généralisation la plus spécifique*.

Nous commençons par emprunter la notation $\wedge$ pour définir une opération sur deux termes t et t' puis nous montrons que $t \wedge t'$ est bien la borne inférieure de t et t'.

DÉFINITION : Opération $\wedge$

On définit $t \wedge t'$ par :
si $f \neq f'$ ou $t = \Omega_S$ ou $t' = \Omega_S$, alors $t \wedge t' = \Omega_S$
si $f = f'$, alors $t \wedge t' = N(f(t_1 \wedge t'_1,..,t_n \wedge t'_n))$

PROPOSITION : $\forall\ t,t' \in \mathcal{T}_S$, $t \wedge t'$ est la borne inférieure de t et t' pour la relation $\leq$. (9)

PREUVE
Il faut montrer que :

•1) $\forall\ t,t' \in \mathcal{T}_S$, $t \wedge t' \leq t$ *et* $t \wedge t' \leq t'$
Ceci est évident quand $t \wedge t' = \Omega_S$.
Par les définitions constructives de $\leq$ et de $t \wedge t'$, on obtient aisément : $f(t_1 \wedge t'_1,..,t_n \wedge t'_n) \leq t$ et $f(t_1 \wedge t'_1,..,t_n \wedge t'_n) \leq t'$
et par la propriété (6) : $N(f(t_1 \wedge t'_1,..,t_n \wedge t'_n)) \leq t$ et $N(f(t_1 \wedge t'_1,..,t_n \wedge t'_n)) \leq t'$

•2) $\forall\ v \in \mathcal{T}_S$, $(v \leq t$ et $v \leq t') \Rightarrow v \leq t \wedge t'$,
Par les définitions constructives de $\leq$ et de $t \wedge t'$, on obtient aisément : $v \leq f(t_1 \wedge t'_1,..,t_n \wedge t'_n)$ et par la propriété (7) on a $N(f(t_1 \wedge t'_1,..,t_n \wedge t'_n))$ est le plus grand terme sous forme normale inférieur à $f(t_1 \wedge t'_1,..,t_n \wedge t'_n)$. Il est donc plus grand que v et on a $v \leq t \wedge t'$. ◁

Exemple 10

L'espace de termes décrivant les piles contient entre autres *pile*(*cube*, *jaune*, *finpile*) et *pile* (*cube*, *vert*, *finpile*). Ces deux termes impliquent l'existence d'un terme *pile*(*cube*, Ω_C, *finpile*) qui est leur borne inférieure et leur généralisation commune la plus spécifique.

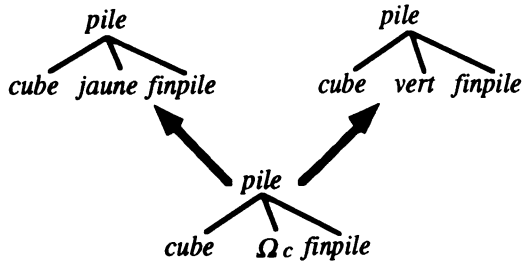

Figure 8

COROLLAIRE : $\mathcal{T}_S$ est un inf-demi-treillis et son minorant universel est Ω_S.

IV.2. Compléter en treillis

$\mathcal{T}_S$ est "presque" un treillis et il est aisé d'en obtenir un en ajoutant un majorant universel.

DÉFINITION : Ajout d'un majorant universel μ

On ajoute un majorant universel μ à $\mathcal{T}_S$, on a : $\forall\, t \in \mathcal{T}_S \cup \{\mu\}$, $t \leq \mu$ et $t \wedge \mu = t$

Pour montrer que $\mathcal{T}_S \cup \{\mu\}$ est un treillis, nous introduisons la notion de compatibilité entre deux termes. Au niveau de l'interprétation, cela consiste à étudier la notion inverse de la généralisation : la spécialisation. Deux termes sont compatibles s'ils possèdent des spécialisations communes c'est-à-dire s'ils généralisent au moins un même terme. Pour cela, il faut que ce qui est spécifié dans l'un corresponde dans l'autre soit à une même spécification, soit à un ou des Ω..

DÉFINITION : Compatibilité

$\forall\, t, t' \in \mathcal{T}_S \cup \{\mu\}$

t et t' sont *compatibles* $\Leftrightarrow$

soit $t = \Omega_S$

soit $t' = \Omega_S$

soit $t = f(t_1,..,t_n)$ et $t' = f(t'_1,..,t'_n)$et $\forall i \in [n]$, t_i et t'_i sont compatibles

μ n'est compatible avec aucun terme.

Exemple 11

L'espace de termes décrivant les piles contient entre autres *pile*(*cube*, Ωc, *finpile*) et *pile* (Ωf,*jaune*, *finpile*). Ces deux termes sont compatibles et impliquent l'existence d'un terme *pile*(*cube*, *jaune*, *finpile*).

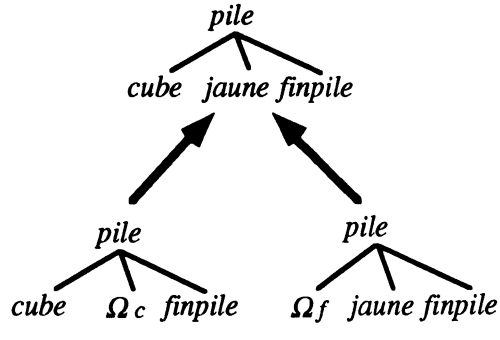

Figure 9

A partir de la compatibilité nous définissons une opération sur deux termes en empruntant la notation $\vee$. Nous démontrons que $t \vee t'$ est bien la borne supérieure de t et t'.

DÉFINITION : Opération $\vee$

Soient $t,t' \in \mathcal{T}_S \cup \{\mu\}$

On définit $t \vee t'$ par :

si t,t' sont compatibles et

si $t = \Omega_S$ alors $t \vee t' = t'$

si $t' = \Omega_S$ alors $t \vee t' = t$

si $t = f(t_1,..,t_n)$ et $t' = f(t'_1,..,t'_n)$ alors $t \vee t' = f(t_1 \vee t'_1,..,t_n \vee t'_n)$

si t,t' ne sont pas compatibles $t \vee t' = \mu$

PROPOSITION : $\forall\, t,t' \in \mathcal{T} \cup \{\mu\}$, $t \vee t'$ est la borne supérieure de t et t' (10)

PREUVE

Il faut montrer que :

• 1) $\forall\, t,t' \in \mathcal{T}_S \cup \{\mu\}$, $t \vee t' \in \mathcal{T}_S \cup \{\mu\}$. C'est-à-dire montrer que si t et t' sont compatibles $t \vee t'$ est sous forme normale (sinon $t \vee t' = \mu$).

Intuitivement, comme aucun Ω n'est ajouté, il n'y a pas de symboles muets qui deviennent inutiles et donc pas de normalisation à effectuer.

• 2) $\forall\, t,t' \in \mathcal{T}_S \cup \{\mu\}$, $t \leq t \vee t'$ et $t' \leq t \vee t'$, ce qui est évident par définition de $\leq$.

• 3) $\forall\, v \in \mathcal{T}_S \cup \{\mu\}$, $(t \leq v$ et $t' \leq v) \Rightarrow t \vee t' \leq v$,

De par les définitions constructives de $\leq$ et de $t \vee t'$, on obtient aisément :

Si t et t' ne sont pas compatibles, le seul majorant commun est μ,

Si t et t' sont compatibles $f(t_1 \vee t'_1,..,t_n \vee t'_n) \leq v$ ◁

COROLLAIRE : $\mathcal{T}_S \cup \{\mu\}$ est un treillis de majorant universel μ.

On peut aisément étendre les définitions des opérations $\wedge$ et $\vee$ et celle de la compatibilité à des sous-ensembles finis de termes. En les considérant comme des algorithmes récursifs (la traduction en Lisp, par exemple est quasiment immédiate), la complexité du calcul est en $O(n \times m)$ où n est la taille de l'ensemble de termes considéré et m le nombre de sommets du plus petit terme de l'ensemble pour la borne inférieure et du plus grand terme pour la borne supérieure.

Notre propos étant de donner une vue globale du modèle, voici une présentation rapide d'autres propriétés structurelles :

On peut montrer que le demi-treillis $\mathcal{T}_S$ vérifie la propriété de Jordan-Dedekind, à savoir : toutes les chaînes couvrantes (chemins maximaux dans la terminologie de la théorie des graphes) entre deux termes sont de même longueur [4]. L'idée est que pour construire un terme t' successeur direct d'un terme t, on ajoute à t un symbole différent de Ω , si ce symbole est muet on génère à partir de lui un sous-terme contenant un seul symbole non muet et différent de Ω . On obtient la propriété : t' est successeur direct de t si son nombre de symboles significatifs (symboles non muets différents de Ω) est supérieur de 1 au nombre de symboles significatifs de

t. . Une chaîne couvrante est une suite de spécialisations directes, Jordan-Dedekind découle donc de cette propriété.

Une autre propriété intéressante est que $\mathcal{T}_S$, quand la signature ne contient pas de symboles muets (dans ce cas $\mathcal{T}_S$ est en fait $\mathcal{T}_S$) est distributif (tous les idéaux principaux sont distributifs). La preuve peut se faire par récurrence sur la hauteur des termes [4]. On peut cependant en donner l'intuition en faisant appel à un raisonnement par analogie : le calcul de la borne inférieure de deux termes ressemble à une "intersection" de leurs sommets et le calcul de la borne supérieure à une "union" (dans le cas où les deux termes sont compatibles). L'intersection et l'union sont distributives l'une par rapport à l'autre.

V. TREILLIS DE GALOIS ET TERMES

Nous proposons dans cette partie une utilisation possible des termes dans le domaine de la classification de données symboliques.

Nous nous sommes intéressés aux travaux de Barbut et Monjardet [1] sur les treillis de Galois. Nous adoptons pour en parler la terminologie de Wille [15] qui a largement contribué à les populariser.

L'idée est de mettre en correspondance deux ensembles partiellement ordonnés. Barbut et Monjardet l'appliquent à partir d'un ensemble d'objets et d'un ensemble d'*attributs* à valeur dans {présent, non-présent}. Plus précisément, un sommet du treillis (ou *concept*) est un couple d'ensembles : le premier est un ensemble d'objets (*définition en extension* du concept), le second un ensemble d'attributs (*définition en intention* du concept) ; tous les objets de l'extension possèdent les attributs de l'intention et celle-ci ne décrit que les objets de l'extension. La relation d'ordre qui induit le treillis est une relation de généralisation : un concept C est plus général qu'un concept C' si l'ensemble d'objets de C inclut celui de C'.

Nous élargissons les domaines réels représentables en admettant des objets structurés. Les descriptions ne sont plus des ensembles d'attributs binaires mais des termes. Nous montrons que nous aboutissons à un treillis de Galois des concepts ainsi généralisés (section V.1)). Nous présentons en seconde section un exemple d'application traitant de la transmission du daltonisme.

V.1. Treillis de Galois et termes

DÉFINITION : t-contexte

Un *t-contexte* est un triplet $(\mathcal{O}, \mathcal{T}_S \cup \{\mu\}, D)$ où

$\mathcal{O}$ est un ensemble fini d'objets

D est une application de $\mathcal{O}$ dans $\mathcal{T}_S \cup \{\mu\}$ qui sert à associer à un objet sa description par un terme sous forme normale (μ ne correspond évidemment à aucun objet).

Pour simplifier, on note, pour $E \subseteq \mathcal{O}$, $D(E) = \{ D(e) / e \in E \}$ et $\wedge D(E) = \bigwedge_{e \in E} D(e)$

DÉFINITION : Applications de la correspondance de Galois

- On définit g, g' deux applications par :

$g : \mathcal{P}(\mathcal{O}) \rightarrow \mathcal{T}_S \cup \{\mu\}$ $\qquad$ $g' : \mathcal{T}_S \cup \{\mu\} \rightarrow \mathcal{P}(\mathcal{O})$

$E \longrightarrow \wedge D(E)$ $\qquad\qquad$ $t \longrightarrow \{o \in \mathcal{O} / t \leq D(o)\}$

avec $g(\varnothing) = \mu$ $\qquad\qquad$ avec $g'(\mu) = \varnothing$

- h est l'application composée g' o g (g' appliquée à g)
  h' est l'application composée g o g'

PROPRIÉTÉS

- g et g' sont monotones décroissantes, c'est-à-dire

$\forall\ E_1, E_2 \in \mathcal{P}(\mathcal{O}),\ E_1 \subseteq E_2 \Rightarrow g(E_2) \leq g(E_1)$ (11)

$\forall\ t_1, t_2 \in \mathcal{T}_S \cup \{\mu\},\ t_1 \leq t_2 \Rightarrow g'(t_2) \subseteq g'(t_1)$ (12)

- h et h' sont monotones croissantes, c'est-à-dire

$\forall\ E_1, E_2 \in \mathcal{P}(\mathcal{O}),\ E_1 \subseteq E_2 \Rightarrow h(E_1) \subseteq h(E_2)$ (13)

$\forall\ t_1, t_2 \in \mathcal{T}_S \cup \{\mu\},\ t_1 \leq t_2 \Rightarrow h'(t_1) \subseteq h'(t_2)$ (14)

- h et h' sont extensives c'est-à-dire

$\forall\ E \in \mathcal{P}(\mathcal{O})\quad E \subseteq h(E)$ (15)

$\forall\ t \in \mathcal{T}_S \cup \{\mu\}\quad t \leq h'(t)$ (16)

PREUVES

- Montrons (11)
  on a $D(E_2) = D(E_1 \cup (E_2 - E_1))$ et $\wedge D(E_1 \cup (E_2 - E_1)) = (\wedge D(E_1)) \wedge (\wedge D(E_2 - E_1))$ par associativité de la borne inférieure $\wedge$ et donc $(\wedge D(E_1)) \wedge (\wedge D(E_2 - E_1)) \leq \wedge D(E_1)$.

- Montrons (12)
  $t_1 \leq t_2 \Rightarrow \forall o \in g'(t_2),\quad t_1 \leq t_2 \leq D(o)$ et donc $o \in g'(t_1)$.

- (13) et (14) sont des conséquences des propositions précédentes.

- Montrons (15)
  $\forall\ o \in E, \wedge D(E) \leq D(o)$ et donc $o \in g'(\wedge D(E))$.
- Montrons (16)
  $\forall\ o \in g'(t), t \leq D(o)$ et donc $t \leq \wedge D(g'(t))$. ◁

Les propriétés de g et g' établissent une correspondance de Galois et on a le théorème suivant :

THÉORÈME

Le couple (g , g') forme une correspondance de Galois entre les ensembles ordonnés $\mathcal{P}(\mathcal{O})$ et $\mathcal{T}_S \cup \{\mu\}$

DÉFINITION : t-concept

Un *t-concept* est un couple (E,t) où $E = g'(t)$ et $t = g(E)$

THÉORÈME

$\mathcal{L}(\mathcal{O}, \mathcal{T}_S \cup \{\mu\}, D)$ l'ensemble des t-concepts de $(\mathcal{O}, \mathcal{T}_S \cup \{\mu\}, D)$ est un treillis de Galois tel que les bornes supérieure ($\vee_G$) et inférieure ($\wedge_G$) de deux t-concepts (E_1,t_1), (E_2,t_2) sont définies par :

$(E_1,t_1) \vee_G (E_2,t_2) = (g'(t_1 \wedge t_2), t_1 \wedge t_2)$

et $(E_1,t_1) \wedge_G (E_2,t_2) = (E_1 \cap E_2, g(E_1 \wedge E_2))$

PREUVE

C'est la définition d'un treillis de Galois appliquée aux treillis $\mathfrak{P}(\mathcal{O})$ muni de l'inclusion et $\mathfrak{N}_S \cup \{\mu\}$ muni de $\leq$. ◁

L'algorithme actuellement implémenté repose sur le fait qu'on peut considérer toutes les branches ("chemins" partant de la racine) des termes comme des attributs (cf. *II.4 termes et variables nominales*). A l'instar de Bordat [3], nous construisons un graphe biparti entre l'ensemble des branches des termes et l'ensemble d'objets. Nous tenons compte du fait que l'ensemble des branches, contrairement à un ensemble d'attributs, possède des propriétés structurelles transmises par les termes. Par une méthode classique basée sur la notion d'intersections ([1]), nous calculons l'ensemble d'objets des t-concepts du treillis puis recomposons les termes à partir de là.

La complexité de l'algorithme est dans le pire des cas en $O(2^n)$ où n = Min (*Nombre total d'objets*, *Nombre de branches*).

V.2. Exemple 12 : application

Nous présentons dans cette section un exemple d'application traitant de la transmission héréditaire du daltonisme.

En France, 1% des hommes sont daltoniens. La transmission génétique est récessive et liée à l'X, ce qui signifie que le caractère daltonien se trouve sur le chromosome X. Un homme le possédant est daltonien, une femme ne l'est que si ses deux X le portent (1 cas sur 10 000 : c'est pourquoi nous ne retenons pas le cas d'une femme daltonienne).
Nous avons construit un échantillon de 101 daltoniens ayant une connaissance partielle du caractère daltonien ou non chez leurs ascendants.

• L'espace des termes sous forme normale décrivant les daltoniens est issu de la signature suivante :
S = {*homme*, *père*, *mère*, *daltonisme*} avec *homme* comme type de base
$F = \{h, p, m, d, n\} \cup \{\Omega_h, \Omega_p, \Omega_m, \Omega_d\}$ (d pour daltonien et n pour non-daltonien)
σ et α sont données dans la table suivante :

F	σ	α	Ω	σ	α
h	*homme*	*père mère*	Ω_h	*homme*	ε
p	*père*	*daltonisme père mère*	Ω_p	*père*	ε
m	*mère*	*père mère*	Ω_m	*mère*	ε
n	*daltonisme*	ε	Ω_d	*daltonisme*	ε
d	*daltonisme*	ε			

On remarque que la récursivité induite par la notion d'ascendance est représentée par les symboles *père* et *mère* et que ceux-ci sont muets.

• Dans notre échantillon, certains daltoniens partagent la même description. L'ensemble considéré peut être partitionné en 7 sous-ensembles $d_1, d_2, d_3, d_4, d_5, d_6, d_7$ chacun correspondant à un terme différent (les effectifs obéissent potentiellement à la réalité, les cas rares étant cependant légèrement surévalués).

$\mathcal{O}$ est donc l'ensemble $\{d_1, d_2, d_3, d_4, d_5, d_6, d_7\}$

Les 36 premiers daltoniens $d_1 = \{1..36\}$ se décrivent par le même terme $t_1 = D(d_1)$ donné dans la figure 10.

t_1 s'interprète par :
- au niveau de l'ascendance paternelle, le père et le grand-père ne sont pas daltoniens, et on ne sait rien pour le reste de l'ascendance.
- au niveau de l'ascendance maternelle, le grand-père n'est pas daltonien, le père de la grand-mère l'est, et on ne sait rien pour le reste de l'ascendance.

De la même façon :
$d_2 = \{37\}$ et $t_2 = D(d_2)$
$d_3 = \{38,..., 47\}$ et $t_3 = D(d_3)$ (en fait $D(d_3)$ est une généralisation du terme $D(d_2)$)
$d_4 = \{48, ..., 67\}$ et $t_4 = D(d_4)$
$d_5 = \{68,69\}$ et $t_5 = D(d_5)$
$d_6 = \{70\}$ et $t_6 = D(d_6)$
$d_7 = \{71, ..., 101\}$ et $t_7 = D(d_7)$

Le treillis de Galois généré est donné dans la figure 11. Par souci de clarté, les sommets ne sont pas tous étiquetés par les t-concepts. La figure 12 développe les sommets où apparaît pour la dernière fois un caractère daltonien et les sommets qui leurs sont supérieurs.

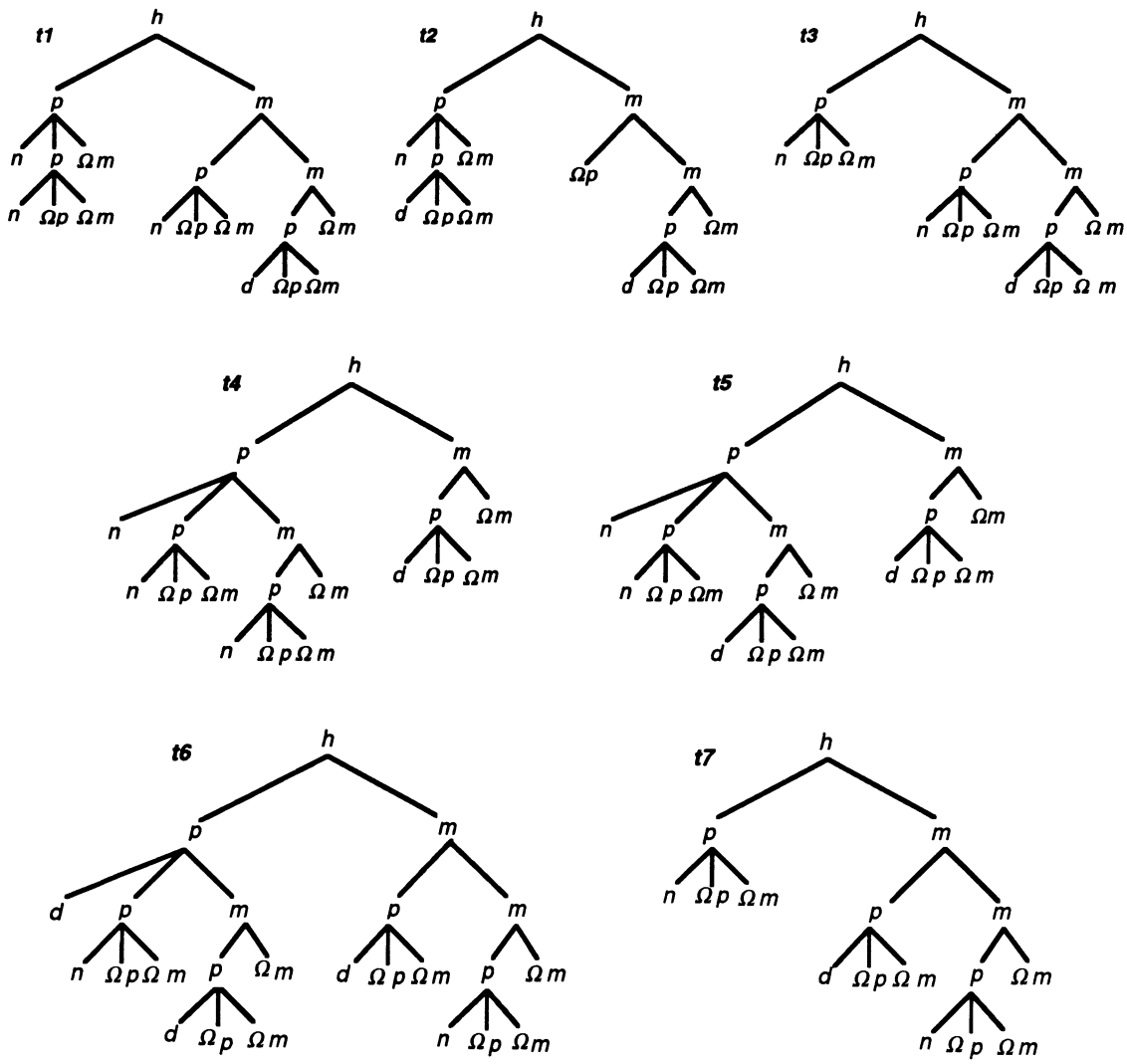

Figure 10

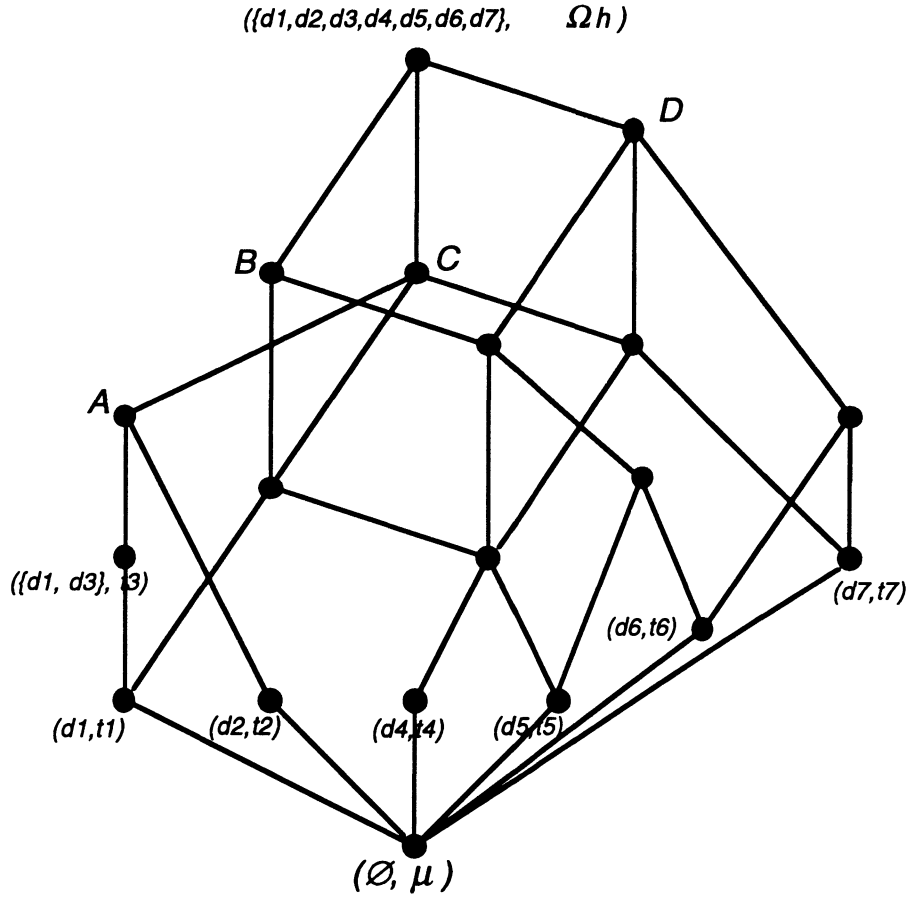

Figure 11

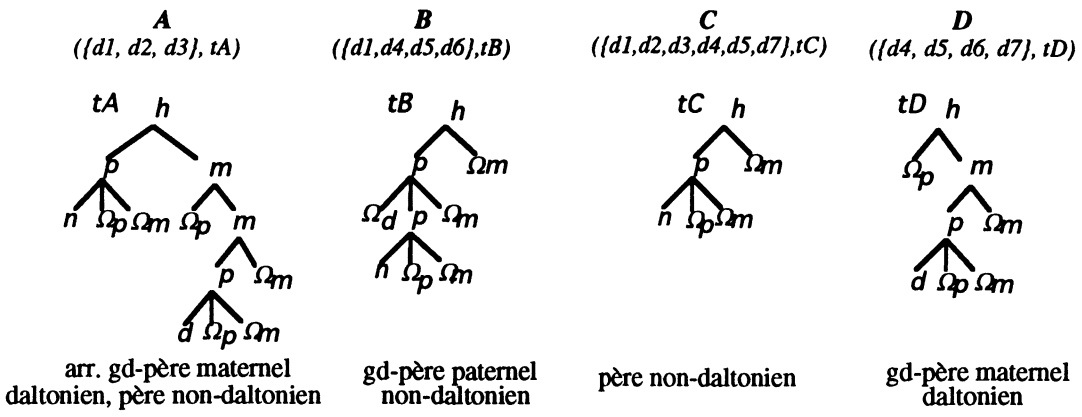

Figure 12

La conclusion est que le daltonisme est dû à l'arrière-grand-père maternel (sommet A décrivant 47 hommes) ou au grand-père maternel (sommet D décrivant 54 hommes) mais pas au père (sommets A, C décrivant 100 pères non-daltoniens) ni au grand-père paternel (sommet B : 60 grand-pères connus non-daltoniens).

Pour une réponse parfaite, le sommet A ne devrait pas mentionner le non-daltonisme du père. Ceci est dû au fait que tous les hommes dont le daltonisme provient de l'arrière-grand-père maternel, partagent aussi la caractéristique d'avoir un père sain.

V. CONCLUSION

Les termes offrent différents avantages :
- leur capacité de description est supérieure à celle des variables nominales. Ils autorisent la représentation de données structurées et en particulier de données dont la nature est récursive,
- leur représentation graphique sous forme d'arborescences permet une interprétation intuitive,
- le rapport existant entre leur sémantique et leur construction syntaxique autorise la présence de symboles d'aide à la visualisation et de symboles entièrement dédiés à leur génération sans pour cela augmenter le nombre de termes et la complexité des traitements puisque qu'on peut ne conserver que les représentants canoniques des classes de termes sémantiquement équivalents,
- leur formalisme est simple, les notions sont définies de façon constructive et le passage aux algorithmes est quasiment immédiat,
- les opérations de base utilisant la relation de généralisation sont polynomiales. On peut ainsi : comparer deux termes, trouver les successeurs ou prédécesseurs immédiats d'un terme, calculer les bornes supérieure et inférieure d'un ensemble de termes.

L'espace d'analyse est le treillis des termes induit par la relation de généralisation. Les bonnes propriétés de cette structure facilitent une exploration partielle (la construction de l'espace entier étant généralement exponentielle voire infinie). Par exemple, la propriété de Jordan-Dedekind permet d'envisager l'utilisation d'algorithmes en largeur d'abord.

Actuellement nous poursuivons l'étude de la structure et cherchons à circonscrire un ensemble de problèmes de caractérisation de classes, polynomiaux pour notre modèle. Notre idée directrice est de définir la recherche de solutions dans le treillis comme une recherche d'antichaînes [13]. En effet, si l'on considère le problème de trouver un ensemble de termes généralisant (caractérisant) un autre ensemble de termes, une "bonne" solution doit au moins assurer que les termes sélectionnés ne sont pas comparables (redondants ou incomplets).

REMERCIEMENTS

Ce travail a été effectué au sein de l'équipe Symbolique-Numérique dirigée par Olivier Gascuel. Les suggestions de celui-ci et ses relectures attentives nous ont permis de l'accomplir.
Merci également à Michel Habib pour ses critiques constructives.

BIBLIOGRAPHIE

[1] BARBUT M., MONJARDET B., *Ordre et classification : Algèbre et combinatoire* (tome I et II), Paris, Collection Hachette Université, Hachette, 1970.

[2] BIRKHOFF G., *Lattice theory*, Colloquium Publications, vol. XXV, AMS, Providence, 3ème édition, 1967.

[3] BORDAT J.P., "Calcul pratique du treillis de Galois d'une correspondance", *Mathématiques et Sciences Humaines*, 24e année, n° 96,1986, pp. 31-47.

[4] DANIEL-VATONNE M.C., *Les termes : un modèle de représentation et structuration de données symboliques*, thèse de Doctorat, Université des Sciences et Techniques du Languedoc, Montpellier II, Février 1993.

[5] DIETTERICH T.G., MICHALSKI R.S., "A Comparating Review of Selected Methods for Learning from Examples", *Machine learning: an Artificial Intelligence approach*, Tioga, Palo Alto 1983, pp. 41-75.

[6] GASCUEL O., "PLage : a way to give and use knowledge in learning", in *Machine and human learning*, Y. Kodratoff Ed., Michaël Horwood series, *Artificial Intelligence*, 1989, pp. 105-120.

[7] GASCUEL O.,GUÉNOCHE A., "Approches symboliques/numériques en apprentissage", *3èmes journées nationales PRC-GDR, Intelligence Artificielle*, B. Bouchon-Meunier Ed., Hermes, 1990, pp. 91-110.

[8] GOGUEN J.A., THATCHER J.W., WAGNER E.G., WRIGHT J.B., "Initial algebra semantics and continuous algebras", *A.C.M*, vol 24:1(1977), pp. 68-95.

[9] GUÉNOCHE A., "Generalization and Conceptual Classification: Indices and algorithm", *Data analysis, learning symbolic and numeric knowledge*, E. Diday Ed., Nova Science Pub 1982, pp. 503-510.

[10] GUESSARIAN I., "Algebraic Semantics", Lecture Notes, *Computer Science* , 99, Goos & Hartmanis Ed., Springer-Verlag, 1981.

[11] HAUSSLER D., "Learning conjunctive concepts in structural domains", *Machine learning*, vol 4 (1989), pp. 7-40.

[12] LIQUIÈRE M., *Apprentissage à partir d'objets structurés*, thèse de Doctorat, Université des Sciences et Techniques du Languedoc, Montpellier II, Février 1990.

[13] MONJARDET B., *Problèmes de transversalité dans les hypergraphes, les ensembles ordonnés, et en théorie de la décision collective.*, thèse de Doctorat d'Etat, Université Paris VI, 1974.

[14] PITT L., REINKE R.E., "Criteria for polynomial-time (conceptual) clustering", *Machine Learning*, vol 2 (1988), pp. 371-396.

[15] WILLE R., "Restructuring lattice theory : an approach based on hierarchies of concepts", *Ordered sets*, I. Rival Ed., Reidel, Dordrecht-Boston,1982, pp. 445-470.

[16] WINSTON P.H., "Learning structural descriptions from examples", *The psychology of Computer Vision*, P.H. Winston Ed., McGraw Hill, New-York, ch.5,1975.