DÉFINITION ABSTRAITE DE LA CLASSIFICATION ET SON APPLICATION AUX TAXONOMIES D'OBJETS

Jérôme Euzenat

INRIA Rhône-Alpes,
LIFIA,
46, avenue Félix Viallet,
F-38031 Grenoble cedex 1
Jerome.Euzenat@imag.fr

RÉSUMÉ

La notion de système classificatoire est introduite comme généralisation de la classification dans les systèmes de représentation de connaissance. Sa définition ne dépend d'aucun modèle de connaissance. Les contraintes qui peuvent lui être ajoutées dans un modèle particulier sont examinées sous la forme de propriétés sémantiques, de structures graphiques et de problèmes d'incomplétude venant entacher les propriétés sémantiques. Ces seules contraintes permettront d'établir certaines propriétés (univocité, déterminance) de l'opération de classification et de concevoir les algorithmes en conséquence. Enfin, le système classificatoire est instancié de deux façons extrêmement différentes dans le cadre du modèle TROPES. La diversité de ces deux dernières interprétations est déjà un exemple de la généralité de cette définition.

MOTS-CLÉS

Objets — Classification — Taxonomie — Catégorisation.

Les acceptions du terme classification, ne serait-ce qu'en intelligence artificielle, sont si nombreuses qu'il est très difficile de comparer les approches sans se référer à un modèle de connaissance particulier. La classification, la taxonomie et la catégorisation sont définies ici en termes relativement généraux totalement indépendants de la façon dont l'appartenance d'un individu à une catégorie est déterminé (§1). Le but de ce travail est d'introduire un modèle de classification qui puisse rendre compte des diverses formes que revêt celle-ci et ainsi de pouvoir:

- établir les propriétés générales des systèmes classificatoires;
- comparer les différentes approches (représentations par objets, types, langages terminologiques,…);
- définir des algorithmes de classification indépendamment de leur instanciation et interprétation, mais en fonction des propriétés générales du système classificatoire sous-jacent.

Trois grandes classes de propriétés sont exposées concernant la structure des systèmes classificatoires (§2 et 3) mais aussi la connaissance des données à classifier (§4) et leurs conséquences sur le processus de classification. Enfin, les systèmes classificatoires sont appliqués à deux modèles très différents: le modèle de connaissance de TROPES qui se rapproche de l'organisation des langages orientés-objets et le modèle de types associé à TROPES qui se rapproche du modèle des langages terminologiques (§5). Cette variété illustre la portée des systèmes classificatoires.

1. SYSTÈME CLASSIFICATOIRE

Un modèle de la classification, indépendant de toute représentation de connaissance, est d'abord proposé. Il définit les notions de catégories (§1.1), de taxonomie (§1.2) et de catégorisation (§1.3). On en trouvera une interprétation au §5.

1.1. Classification

La classification agit sur des ensembles d'entités abstraites L_C et L_I dont les éléments sont nommés respectivement catégories et individus. Un individu i peut être classé sous une catégorie c. Ce que sont ces entités ne sera pas plus développé; elles trouvent des interprétations diverses suivant les modèles de connaissance considérés. Le classement d'un individu dans une catégorie peut être considéré comme le collage d'une étiquette sur celui-ci.

À partir d'un ensemble $C(\subseteq L_C)$ de catégories, l'opération de classification détermine l'ensemble Cl(i,C) de catégories dans lesquelles un individu particulier i est classé (dans la suite la notation Cl(i) sera utilisée car elle n'est pas ambiguë). L'ensemble des individus qui doivent être attachés à une catégorie c est nommé $In(c) = \{i \in L_I ; c \in Cl(i)\}$.

1.2. Taxonomie

L'ensemble des catégories est habituellement organisé en une taxonomie (dans le sens utilisé dans [WOO91], par exemple). Une taxonomie est une structure faite d'un ensemble C de catégories et d'une relation d'ordre nommée sous-catégorisation ($\leq$) sur ces catégories telle que: $c' \leq c \Rightarrow In(c') \subseteq In(c)$. Dans la suite, $\leq$ sera confondue avec son graphe sur C. La taxonomie est utile en tant que représentation d'une propriété abstraite des catégories: celle d'aller des plus

générales aux plus spécifiques. Elle peut être utilisée pour implémenter des algorithmes de classification efficaces.

Il est pratique de représenter le graphe d'une relation d'ordre sous la forme de sa réduction transitive et réflexive [AHO72]. De même, il est plus simple de représenter le résultat de la classification d'un individu sous la forme de sa minimisation par rapport à la relation de sous-catégorisation. La minimisation d'un ensemble S par rapport à un ordre partiel R sur ses éléments est définie par $\mu_R S = \{x \in S; \forall y \in S, yRx \Rightarrow xRy\}$: elle ne retient que les éléments de S minimaux pour R. Les catégories minimales pour un individu i sont donc $\mu_{\leq} Cl(i)$. Les catégories terminales d'un individu i sont ses catégories qui n'ont pas de sous-catégories (c'est-à-dire celles qui sont des puits du graphe de sous-catégorisation): $\nu Cl(i) = Cl(i) \cap \mu_{\leq} C$. L'ensemble $\mu_{\leq} C$ est l'ensemble des puits de la taxonomie.

1.3. Catégorisation

La catégorisation (aussi appelée classification de classes) permet la construction d'une taxonomie. Elle détermine incrémentalement les relations entre une nouvelle catégorie et celles qui sont déjà dans la taxonomie. La catégorisation construit donc incrémentalement la relation ≤. Ceci est réalisé à l'aide d'un critère de sous-catégorisation (noté <<) défini sur L_C. << est différent de ≤, défini sur C, qui reste à construire. Bien souvent, le critère de sous-catégorisation est bâti à partir d'un ordre partiel sur L_C (par exemple, relation de sous-typage, d'inclusion,…). Intuitivement, il est la contrainte la plus lâche que doit respecter la relation de sous-catégorisation: une catégorie peut être insérée n'importe où dans le graphe de sous-catégorisation tant que la relation obtenue respecte le critère de sous-catégorisation:

$$\forall c, c' \in C, c \leq c' \Rightarrow c << c'$$

qui sera tel que:

$$\forall c, c' \in L_C, c << c' \Rightarrow In(c) \subseteq In(c')$$

Pour que l'insertion d'une nouvelle catégorie respecte ce critère, il faudra qu'elle soit sous-catégorie de catégories plus générales au sens de << et que les catégories qui seront ses sous catégories soient plus spécifiques au sens de <<. Ainsi, l'opération de catégorisation établit-elle, pour une catégorie c, l'ensemble de ses catégories les plus générales ($MGC(c) = \{c' \in C; c << c'\}$) et l'ensemble de ses catégories les plus spécifiques ($MGC(c) = \{c' \in C; c' << c\}$) par rapport à <<. Ces deux ensembles sont plus facilement représentés par les catégories plus générales les plus spécifiques ($MSMGC(c) = \mu_{\leq} MGC(c)$) et les catégories plus spécifiques les plus générales ($MGMSC(c) = \mu_{\geq} MSC(c)$) respectivement. À noter que la minimisation est effectuée en fonction de ≤ qui est déjà établie sur C: cette minimisation est pertinente pour le graphe à construire.

La catégorisation telle qu'elle est présentée ici ne couvre pas la notion de classification (construction de classification hiérarchisée) en analyse de données ou en apprentissage symbolique automatique. En effet, il s'agit ici d'inférence des positions possibles d'une catégorie déjà définie au sein d'une taxonomie alors que la construction de classification hiérarchisée infère les définitions de catégories (et les liens qu'elle entretient avec les autres catégories) en fonction d'un ensemble donné d'individus. Les travaux sont en cours pour établir une définition de la construction de classification hiérarchisée dans le cadre des systèmes classificatoires.

Ainsi donc, un système classificatoire sur un couple de langages L_C et L_I est donné par un ensemble de catégories (C), une relation de sous-catégorisation ($\leq$) et un critère de sous-catégorisation ($<<$). Le premier permet de définir la classification, la seconde, la taxonomie et le dernier, la catégorisation. La taxonomie ($\leq$) est un appauvrissement de la restriction à C de la relation $<<$ définie sur L_C. Bien entendu, cet appauvrissement peut ne pas avoir lieu, auquel cas $\leq$ est la simple restriction à C de $<<$ et l'ajout d'une catégorie à C est immédiat.

Cette définition de la classification est très générale: l'interprétation des notions utilisées n'est pas fournie et les contraintes pesant sur chacun des composants sont faibles. Dans les sections 2, 3 et 4, des contraintes qui peuvent être introduites, en particulier sur la relation de sous-catégorisation et sur les individus, vont être présentées afin d'en étudier les conséquences.

2. PROPRIÉTÉS SÉMANTIQUES DE LA TAXONOMIE

Les propriétés sémantiques concernent le comportement de la taxonomie lors de la classification par opposition à ses propriétés purement graphiques. En particulier, pour garantir que la classification retournera toujours un unique résultat minimal, il est possible de poser deux contraintes:

- Interdire au processus de classification de diverger localement (de considérer deux catégories incomparables par $\leq$ comme appartenant à Cl(i)).
- Imposer que le processus de classification converge globalement (autoriser à diverger localement sachant que les alternatives mèneront à un minimum).

2.1. Exhaustivité et exclusivité

Une propriété essentielle est l'exhaustivité: chaque individu classé dans une catégorie est classé dans au moins une de ses sous-catégories:

$$\forall c,(\exists c';c' \leq c \wedge c' \neq c) \Rightarrow (\forall i,c \in Cl(i) \Rightarrow \exists c';c' \leq c \wedge c' \neq c \wedge c' \in Cl(i))$$

Cette propriété traduit l'intention de certaines taxonomies en biologie dans lesquelles tout individu fait nécessairement partie d'un taxon terminal (il n'existe pas de purs mammifères, c'est-à-dire de mammifères qui ne soient pas membres d'une sous-catégorie (ordre): ce sont soit des primates, soit des rongeurs,…). Un des intérêts de l'exhaustivité est que si un individu est classé dans une catégorie, mais dans aucune de ses sous-catégories sauf une, alors l'individu doit être classé dans cette dernière. Plus intéressant encore, la transitivité permet de dire que si un individu est classé dans une catégorie, mais dans aucune de ses sous-catégories directes (dans la réduction transitive du graphe de $\leq$), alors il doit être classé dans cette dernière.

Une autre propriété importante est l'exclusivité (préférée ici à «disjonction» parce qu'elle représente la disjonction exclusive et non la disjonction logique): l'ensemble des catégories dans lesquelles un individu peut être classé forme une chaîne pour $\leq$.

$$\forall c',c'',c'' \leq c' \vee c' \leq c'' \vee \forall i,\neg(c' \in Cl(i) \wedge c'' \in Cl(i))$$

ou encore

$$\forall c',c'',(\exists i;c' \in Cl(i) \wedge c'' \in Cl(i)) \Rightarrow (c' \leq c'' \vee c'' \leq c').$$

Cette exclusivité peut aussi s'exprimer sous une forme réduite, plus locale, qui stipule qu'un individu classé dans une catégorie ne peut être classé que dans une de ses sous-catégories directes (c'est-à-dire dans la réduction transitive de $\leq$). Les deux expressions sont équivalentes.

2.2.Univocité et déterminance

L'univocité signifie qu'il existe toujours un minorant à deux catégories dans Cl(i):

$$\forall i, \forall c', c'' \in Cl(i), \exists c \in Cl(i); c \leq c' \wedge c \leq c''$$

si Cl(i) est un ensemble fini. Si tel n'était pas le cas la propriété se réécrirait:

$$\forall i, \exists c; \mu_{\leq} Cl(i) = \{c\} \text{ ou } \forall i, |\mu_{\leq} Cl(i)| = 1.$$

Elle est importante car si $|\mu_{\leq} Cl(i)|=1$, alors l'individu i a été classé de manière unique: cela entraîne de nombreuses conséquences sur les algorithmes et les interprétations possibles. En particulier, lorsqu'un individu ne peut être effectivement affecté (par la représentation de connaissance pour le compte de laquelle est réalisée la classification) qu'à une seule catégorie minimale, l'univocité est nécessaire à la classification automatique (c'est-à-dire l'affectation d'une catégorie à un individu en fonction du résultat de la classification). À noter que cette propriété est plus générale que la propriété d'exclusivité (au sens où si la propriété d'exclusivité est satisfaite, celle-ci l'est aussi). En particulier, elle s'applique à des structures de treillis ce que l'exclusivité interdit.

Par ailleurs si $\mu_{\leq} Cl(i) = \nu Cl(i)$, alors l'individu a été complètement identifié eu égard à la taxonomie; la classification est alors déterminante. Il est donc important de disposer de systèmes classificatoires répondant à ces caractéristiques. Cela est si contraignant qu'il est douteux que de tels systèmes soient courants.

2.3.Propriétés

Les propriétés sémantiques, lorsqu'elles sont accessibles à l'opération de classification (voir §4.1), induisent des propriétés très importantes sur ses résultats:

- Si la relation de sous-catégorisation est exclusive, le système classificatoire est univoque (il n'y a qu'une seule solution minimale: $\forall i, |\mu_{\leq} Cl(i)| = 1$).
- Si la relation de sous-catégorisation est exhaustive, le système classificatoire est déterminant (les solutions minimales sont aussi terminales: $\forall i, \mu_{\leq} Cl(i) = \nu Cl(i)$).

Quand la relation de spécialisation est exclusive et exhaustive, il n'y a qu'une seule solution minimale qui est aussi terminale ($\forall i, \exists c; \mu_{\leq} Cl(i) = \nu Cl(i) = \{c\}$). Ces deux propriétés sont celles qui sont exigées d'une taxonomie telle que celles utilisées en biologie, par exemple.

3.STRUCTURE GRAPHIQUE DE LA TAXONOMIE

Le graphe de la relation de sous-catégorisation peut être étudié car, en tant que support de la minimisation, il contraint les propriétés possibles de la classification. Ce faisant, il ne permet que des propriétés sémantiques sans les imposer. Mais la structure graphique de la taxonomie est révélatrice des options utilisées dans le système classificatoire. Par ailleurs, il définit les chemins potentiels que peut emprunter un algorithme de classification.

3.1.Demi-treillis inférieur

La structure minimale exigée de $<C,\leq>$ est celle d'un ordre partiel (réflexif, transitif et anti-symétrique). Cela peut être obtenu à partir de n'importe quel pré-ordre par passage de C au quotient modulo $\leq$. Le graphe correspondant à la relation est par conséquent un graphe orienté sans circuit (DAG). Une structure

pertinente pour la classification est celle de semi-treillis inférieur. En introduisant l'obligation de disposer d'une plus grande borne inférieure (glb, notée ∇) sur tout couple de catégories, elle permet de supporter l'univocité[1]:

$$\forall i, \forall c', c'' \in Cl(i), c' \nabla c'' \in Cl(i).$$

Ceci n'est en effet possible que lorsque c'∇c" est défini, c'est-à-dire lorsque la plus grande borne inférieure existe.

À noter que la condition $(\forall i, \forall c', c'' \in Cl(i), c' \nabla c'' \in Cl(i))$ est nécessaire pour que la structure de treillis agisse pleinement. Par exemple, si c' est «professeur de musique», c" «professeur avec un chat» et c «professeur de musique avec un chat et trois enfants», il est vrai que $c \leq c'$ et $c \leq c''$. Cependant, un professeur de musique avec un chat et deux enfants sera attaché à c' et c" mais pas à c.

La structure de semi-treillis inférieur interdit la propriété d'exclusivité (sauf si le graphe est réduit à une chaîne). Un semi-treillis inférieur possédant la propriété d'exhaustivité conduit à la propriété que Cl(i) est réduit au singleton contenant l'élément minimal du treillis, et ceci quelque soit i.

3.2. Arbre recouvrant

La relation $\leq$ étant transitive, le graphe correspondant ne pourra être un arbre (graphe connexe avec une source et tous les autres nœuds de degré rentrant égal à 1) ou une forêt (ensemble d'arbres) que si tout élément du graphe est soit un puits, soit une source, ce qui est de fort peu d'intérêt. Par contre, la structure de la réduction transitive et réflexive de $\leq$ peut être un arbre. Disposer d'un arbre semble aller à l'encontre des préoccupations de la section précédente: en effet, celui-ci n'admet plus de borne inférieure entre deux éléments que lorsque ceux-ci sont comparables.

L'arbre est le support naturel de la propriété d'exclusivité, car si cette dernière est vérifiée dans un graphe orienté sans circuit, alors toutes les catégories de degré rentrant supérieur à 2 ne peuvent être dans aucun résultat de classification.

Ainsi, les structures graphiques de la taxonomie ne portent, en elles-mêmes, aucun sens, mais elles sont révélatrices de ce qu'il peut advenir lors de la classification. À cet égard, elles sont de bons indices des propriétés sémantiques associées aux taxonomies: lorsque l'exclusivité est requise, il est judicieux de disposer d'un arbre (pour la réduction de l'ordre); lorsque la convergence est exigée, il est souhaitable de disposer d'un ∇-semi-treillis. Si tel n'est pas le cas, alors, au mieux, certaines potentialités de la taxonomie seront inutilisées.

4. INCOMPLÉTUDE

Les sections précédentes ont considéré que les informations concernant la taxonomie et les individus étaient complètes et accessibles au classificateur. En fait, comme souvent, toutes ces informations ne sont pas forcément connues et le système de classification doit en tenir compte.

---

[1] Ce n'est pas suffisant dans le cas où Cl(i) peut être infini: il faut en plus que le treillis soit complet, c'est-à-dire que $\forall S \subseteq C, |\mu_{\leq}\{c \in C; \forall c' \in S, c \leq c'\}| = 1$

4.1. Subjectivité et objectivité

Exclusivité et exhaustivité peuvent être affectées d'un caractère objectif ou subjectif. Intuitivement, une propriété est subjective lorsqu'elle est dans le sujet (celui qui construit la taxonomie ici) et objective lorsqu'elle est dans l'objet (l'énoncé de la catégorie ici). Une propriété sera définie comme objective si elle peut être déduite et subjective si elle doit être valide mais ne peut être déduite. Par exemple, dans les représentations par objets, les attachements procéduraux ne doivent pas avoir d'effets de bord. Cette propriété est subjective car le système ne peut la vérifier.

Ce caractère est lié à l'incomplétude des bases de connaissance: elles ne peuvent renfermer toutes les données qui permettraient d'établir objectivement les propriétés. Savoir qu'une propriété doit être vérifiée, même subjectivement, peut être très utile: s'il est prouvable que la propriété n'est pas vérifiée, il y a un problème que le système peut rapporter à l'utilisateur. Par exemple, si l'exhaustivité subjective est requise d'un système classificatoire et qu'un algorithme de classification correct trouve qu'un individu i est classé dans une catégorie mais dans aucune de ses sous-catégories, il peut signaler une violation de la propriété d'exhaustivité.

Lorsque les propriétés sémantiques ne sont que subjectives, le système ne peut en tirer parti et les conséquences qui découlent de ces propriétés ne sont plus vérifiées: il faut donc en tenir compte dans les algorithmes. Par exemple, si l'exclusivité est prouvable par le système, alors un algorithme de classification peut parcourir le graphe en profondeur et s'arrêter à la première catégorie dont aucune sous-catégorie ne permet de classer l'individu. Si l'exclusivité n'est que subjective, ceci est périlleux.

4.2. Objets incomplets

La connaissance de l'individu i à classer peut être incomplète, ce qui signifie que l'opération de classification sera incapable de dire si une catégorie c appartient à Cl(i) ou si au contraire elle n'y appartient pas. Pour exprimer cela, deux ensembles sont utilisés: $Cl_p(i)$, l'ensemble des catégories dont on sait qu'elles appartiennent à Cl(i), et $Cl_u(i)$, celles dont on ignore si elles appartiennent ou non à Cl(i). Plus formellement: $Cl_p(i) \subseteq Cl(i)$ et $Cl_u(i) \subseteq C \backslash Cl_p(i)$.

Soit un animal à classer dans une taxonomie animale, la connaissance de son squelette permet de la faire entrer dans la catégorie des vertébrés. Mais l'absence de connaissance sur le revêtement de sa peau interdit de le classer dans les oiseaux ou les reptiles[2]. Pourtant la taxonomie en question est exhaustive et:

(1) l'individu en question appartient bien à l'une des deux catégories;
(2) la taxonomie garantit effectivement que l'individu appartient à une catégorie terminale.

Cependant, le manque de connaissance sur l'individu ne permet pas d'en dire plus. Ainsi, la catégorie des vertébrés appartient bien à Cl(i), elle est dans $Cl_p(i)$, les catégories reptiles et oiseaux appartiennent uniquement à $Cl_u(i)$ — pour être complet, on sait que l'une d'elles appartient à Cl(i) —, et la catégorie des mammifères n'appartient ni à l'un, ni à l'autre.

Les résultats mentionnés ci-dessus en fonction de l'exclusivité et de l'exhaustivité ne sont plus valides pour ces ensembles. En particulier, si $\forall i, \mu_{\leq} Cl(i) = \nu Cl(i)$, ceci n'est plus vrai pour $Cl_p(i)$. En effet, s'il existe un individu

[2] Bien entendu, des informations génétiques permettront d'identifier l'animal.

dont on ne sait pas s'il appartient ou non à une catégorie c (donc si $c \in Cl(i)$), alors on ne sait pas non plus s'il appartient à chacune des sous-catégories de c. Ainsi, $\mu_{\leq}Cl_p(i)$ contient des catégories non terminales. De même, si $|\mu_{\leq}Cl_p(i)|=1$ est vrai pour un individu i mais pas généralement prouvable, ceci n'est plus forcément vrai de $Cl_p(i)$ car il se peut que cet individu soit classé dans deux catégories minimales différentes alors qu'il existe une catégorie minimale plus spécifiques que ces deux là.

Il faut noter que la structure de treillis est très robuste vis-à-vis de l'incomplétude. En effet, la connaissance incomplète d'un individu à classer dans un treillis entraîne bien deux ensembles $Cl_p(i)$ et $Cl_u(i)$, mais la minimisation du premier reste un singleton. Ceci se déduit de la propriété même du treillis (si plusieurs catégories sont connues pour être dans Cl(i), ceci est aussi vrai de leur plus grand minorant).

5. APPLICATION À LA CLASSIFICATION ET CATÉGORISATION DANS TROPES

La définition de la classification et ses possibles propriétés données ci-dessus ne signifient rien hors de l'interprétation de ses résultats. Les systèmes de représentation de connaissance utilisant la classification sont fort nombreux et interprètent diversement la classification et la catégorisation. Ces différentes interprétations ne seront pas détaillées ici (voir [NAP92] par exemple), mais l'application du présent travail au système TROPES permet déjà d'en donner deux interprétations très différentes et sera donc évoquée.

5.1. Rapide aperçu de TROPES

TROPES [MAR90] est un modèle de représentation de connaissance par objets. Les individus sont donc les objets. Ceux-ci sont partitionnés en concepts: un objet est instance d'un unique concept. Les concepts et leurs instances sont visibles sous divers points de vue. Un point de vue détermine une hiérarchie de classes (c'est-à-dire un ensemble de classes structuré par une relation nommée spécialisation). Dans chaque point de vue, un objet est attaché à (par opposition à «est instance de») une unique classe plus spécialisée (sachant qu'il est alors attaché à toutes les classes dont celle-ci est spécialisation). Chaque point de vue offre à l'utilisateur de la classification une taxonomie différente dans laquelle le classement dans une classe dépend de critères différents. Ils permettent de se focaliser sur certains aspects des objets sans être gêné par les autres.

Les concepts déterminent les attributs possibles des instances et le type de leurs valeurs. Les classes déterminent parmi ceux-ci les attributs pertinents des objets qui leur sont rattachés et raffinent le type de leurs valeurs. À une classe peut être associé un type d'enregistrement ("record") dans l'esprit de [CAR91]. Il faut noter que le fait qu'un objet soit compatible avec le type associé à une classe n'implique pas qu'il doive être attaché à celle-ci. En effet, l'interprétation des classes est descriptive, c'est-à-dire que la compatibilité avec le type d'une classe est une condition nécessaire mais non suffisante pour le rattachement d'un objet à cette classe. La condition suffisante est l'acte volontaire de rattachement effectué par l'utilisateur du système ou un programme d'application. Pour la défense d'une sémantique descriptive des classes, il est possible d'évoquer la généralité de l'interprétation qui autorise toujours à la restreindre localement lorsque cela est possible et permet un plus large champ d'applicabilité au système.

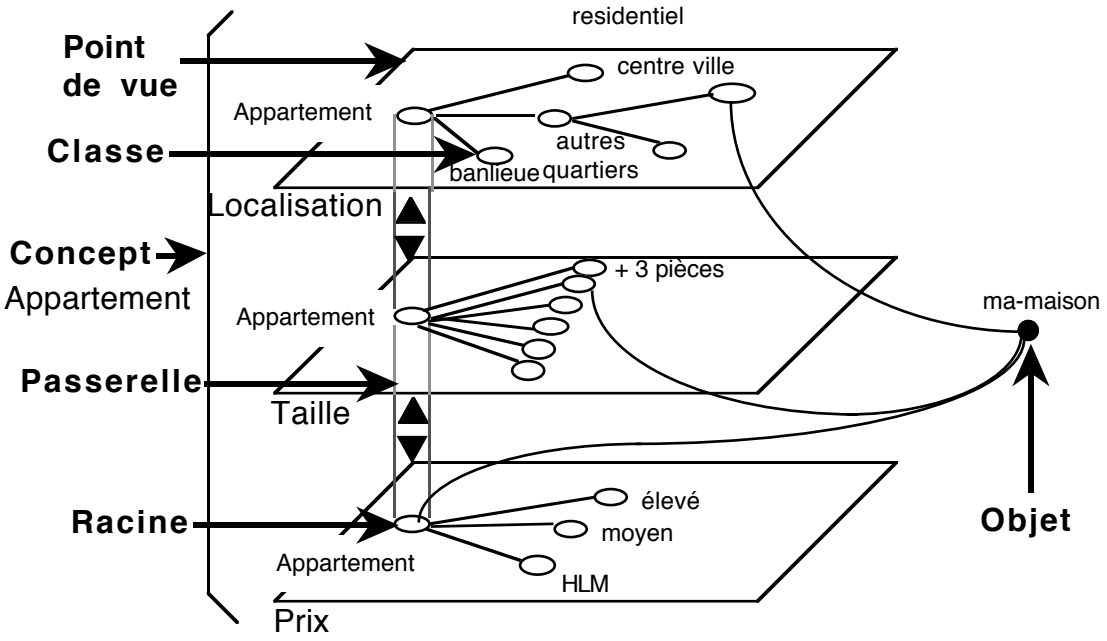

Figure: Le concept Appartement est visible sous les points de vue Localisation, Taille et Prix. Chaque point de vue détermine une taxonomie de classes dont la racine se nomme Appartement. Par exemple, sous le point de vue Localisation, il y a une décomposition des classes suivant les quartiers de la ville. L'objet ma-maison est attaché sous le point de vue Localisation à la classe Résidentiel, sous le point de vue Taille à Plus de 3 pièces et sous le point de vue Prix à Élevé.

La relation de spécialisation dans TROPES a deux interprétations: elle peut être considérée comme la relation d'inclusion ($\subseteq$) entre les ensembles d'objets attachés aux classes ou comme la relation de sous-typage entre des types associés aux classes (<:). Chacune de ces interprétations est correcte mais non complète.

5.2.Classification et catégorisation dans TROPES

TROPES définit des systèmes classificatoires sur les langages de classes et d'objets dont l'ensemble des catégories correspond aux classes d'un point de vue, la relation de sous-catégorisation est la (clôture réflexive de la) relation de spécialisation et le critère de sous-catégorisation est la relation de sous-typage (<:) que doit respecter la spécialisation.

Les propriétés de la relation de spécialisation sont les suivantes:

- Sa réduction transitive est un arbre;
- Elle est non exhaustive;
- Elle est subjectivement exclusive;

Par ailleurs, les instances peuvent être incomplètes.

Suite aux résultats énoncés dans les sections précédentes, l'ensemble $\mu_{\leq}\mathrm{Cl}(i)$ n'est pas nécessairement un singleton, n'est pas vide (puisque la classe initiale est dans Cl(i)) et n'est pas forcément un ensemble de classes terminales. La conséquence immédiate en est que le système classificatoire associé n'est pas univoque et donc la classification automatique est impossible dans TROPES. Ce résultat n'est que pure formalité puisque la sémantique descriptive des classes à elle seule interdit la classification automatique (même si le graphe de la relation était objectivement exhaustif et exclusif).

Cependant, dire que la classification automatique est impossible n'interdit pas toute interprétation de l'opération de classification. Cette interprétation, en accord

avec la sémantique descriptive des classes, est qu'un objet peut être attaché à n'importe quelle classe parmi celles dans lesquelles il a été classé. Mais c'est à l'utilisateur de fournir la condition suffisante à ce rattachement par une action d'attachement. Similairement, la catégorisation dans TROPES signale que la classe catégorisée peut être spécialisation de n'importe laquelle de ses classes les plus générales et que n'importe lesquelles de ses classes les plus spécifiques (qui sont aussi plus spécifiques que la classe choisie précédemment) peut être de ses spécialisations. Encore une fois, c'est l'utilisateur qui choisira les classes concernées.

L'algorithme de classification subit donc la sémantique énoncée: l'objet i est nécessairement attaché à une classe initiale c (qui peut être la racine de la taxonomie); ses valeurs d'attributs sont nécessairement conformes au type de cette classe initiale (même si la valeur en est inconnue). Il respecte donc les conditions nécessaires et suffisantes pour être classé sous cette classe c. L'algorithme descend dans les sous-classes testant la conformité de l'objet au type associé à la classe. La relation de sous-type existant entre types associés aux classes liées par la spécialisation permet d'épargner certains tests à l'algorithme et d'en déduire, à partir d'une classe n'appartenant pas à Cl(i), que toutes ses sous-classes en sont aussi exclues [HAT91].

5.3.Catégorisation dans le treillis de types associés

Il a été proposé [CAP93] d'utiliser le langage exprimant les types associés aux classes pour assurer la catégorisation. Chaque classe se voit associer un type unique; mais à un type peut correspondre aucune, une ou plusieurs classes. Ce langage doit avoir certaines propriétés comme celle de normalisation (deux expressions dénotant le même type sont syntaxiquement égales). L'ensemble des types constructibles à partir d'un tel langage structuré par la relation de sous-typage (<:) forme un treillis complet. Le système n'entretient pas l'ensemble du treillis des types qui est souvent infini. Il se contente de maintenir le sous-treillis minimal contenant les types correspondant à des classes.

Le système classificatoire correspondant est donc constitué de l'ensemble des types utilisés en tant qu'ensemble de catégories, la relation de sous-typage comme relation de sous-catégorisation et la même relation de sous-typage comme critère de sous-catégorisation. Il est univoque.

Ainsi, lorsqu'il s'agit de classer un objet, le système peut automatiquement retrouver son type le plus spécifique (qui est unique). Le classement d'une valeur dans le treillis des types peut donc être réalisé automatiquement. Mais lorsqu'il s'agit d'attacher l'objet correspondant à cette valeur à une classe, il peut y avoir plusieurs candidates: toutes les classes dont le type correspondant est plus général que le type le plus spécifique calculé. De même, trouver la position d'un nouveau type dans le treillis consiste à trouver l'ensemble de ses sur-types et l'ensemble de ses sous-types. Une fois ceci fait, le type est automatiquement introduit dans le treillis puisqu'il est effectivement sous-type de tous ses types plus généraux et sur-type de tous ses types plus spécifiques. La catégorisation automatique est donc aussi possible dans le treillis des types (il faut ensuite mettre à jour le graphe pour conserver la structure de treillis). Mais lorsqu'il s'agit d'insérer la classe associée au type dans le graphe de spécialisation, ceci ne peut toujours pas être automatisé: l'utilisateur doit choisir la sur-classe parmi les classes correspondant aux types sur-types du type courant et les sous-classes parmi les spécialisations directes de la classe choisie. Ce procédé, utilisant directement les types au lieu de passer par la relation de spécialisation, a des chances d'accélérer la catégorisation mais peut aussi gêner l'utilisateur.

Dans un système disposant de multi-spécialisation (la réduction transitive de la relation de spécialisation est alors un graphe orienté sans circuit et non un arbre, mais la sémantique de la relation de spécialisation demeure la même), il est possible d'automatiser la construction de la taxonomie. Il suffit de considérer que chaque classe est spécialisation des classes correspondantes aux sur-types du type qui lui est associé et que les classes correspondantes aux sous-types du type qui lui est associé sont ses spécialisations. Si, de surcroît, à un type ne correspond qu'une classe (ou plutôt une classe d'équivalence entre classes), alors le graphe de spécialisation est isomorphe au sous-treillis des types et le système classificatoire est univoque. Ceci est semblable à ce qui est réalisé dans les langages terminologiques [BRA83].

6. CONCLUSION

Après avoir défini la classification dans un cadre extrêmement général, les systèmes classificatoires, les propriétés permettant d'obtenir des résultats algorithmiques sur la classification ont été proposées. Le rapport entre ces propriétés et les structures graphiques des taxonomies a été abordé jusqu'à en montrer les limites (incomplétude des représentations, nécessité de critères sémantiques). Toutes ces informations sont nécessaires à la caractérisation de la classification dans une taxonomie d'objets.

Deux interprétations de la classification telle qu'elle est définie ici ont aussi été produites. Il est remarquable que la caractérisation de la classification permette d'intégrer des mécanismes aussi différents que la classification dans un arbre de classes descriptives et que l'inférence de relations de sous-typage. On se convaincra aisément que le schéma présenté ici s'applique aussi aux langages terminologiques et aux arbres de décision. Il reste à établir si ce schéma s'applique toujours dans le cas de classifications fondées sur des mesures de similarités numériques ou de probabilités. Il est d'envisageable de graduer les réponses aux opérations de classification et de catégorisation (par exemple, faire de Cl(i) un ensemble flou). La comparaison avec de tels classificateurs pourrait conduire à généraliser encore la notion de classification proposée ici.

REMERCIEMENTS

Que soient remerciés ici les lecteurs anonymes ainsi que Cécile Capponi, Mathias Chaillot, Jérôme Gensel et Patrice Uvietta pour leurs lectures précises et attentives. Tout problème résiduel demeure de mon fait.

RÉFÉRENCES

[AHO72] Alfred Aho, Michael Garey, Jeffrey Ullman, **The transitive reduction of a directed graph**, SIAM journal of computing 1(2):131-137, 1972

[BRA83] Ronald Brachman, Richard Fikes, Hector Levesque, **KRYPTON: a functional approach to knowledge representation**, IEEE Computer 16(10):30-36, 1983

[CAP93] Cécile Capponi, **Classification des classes par les types**, Actes 2ndes journées représentation par objets, La Grande Motte (FR), ce volume, 1993

[CAR91] Luca Cardelli, John Mitchell, **Operations on records**, Mathematical structures in computer science 1(1):3-48, 1991

[HAT91] Jean-Paul Haton, Nadjet Bouzid, François Charpillet, Marie-Christine Haton, Brigitte Lâasri, Hassan Lâasri, Pierre Marquis, Thierry Mondot, Amedeo Napoli, **Le raisonnement en intelligence artificielle**, InterÉditions, Paris (FR), 1991

[MAR90] Olga Mariño, François Rechenmann, Patrice Uvietta, **Multiple perspectives and classification mechanism in object-oriented representation**, Actes 9th ECAI, Stockholm (SE), pp425-430, 1990

[NAP92] Amedeo Napoli, **Représentations à objets et raisonnement par classification en intelligence artificielle**, Thèse d'état, Université de Nancy 1, Nancy (FR), 1992

[WOO91] William Woods, **Understanding subsumption and taxonomy: a framework for progress**, dans John Sowa (éd.), Principle of semantic networks: exploration in the representation of knowledge, pp45-94, Morgan-Kauffman, San-Mateo (CA US), 1991